HEIGHT APPROXIMATION FOR AUDIO SOURCE LOCALISATION AND TRACKING

Ashley Hughes[⋆] *James R. Hopgood*[⋆] *Neil M. Robertson*[†]

[⋆]School of Engineering, The University of Edinburgh [†]Visionlab, Heriot-Watt University, Edinburgh
Email: {a.hughes, james.hopgood}@ed.ac.uk n.m.robertson@hw.ac.uk

ABSTRACT

The stochastic region contraction (SRC) algorithm has been proposed in the literature as a method for acoustic localisation using a microphone array in a noisy and reverberant environment. This technique makes use of the steered response power (SRP), a costly but robust technique for source localisation, and finds the global maximum vastly more efficiently than using a grid search method. We discuss combining this technique with prior information (e.g. in future work we will use a video tracker) to speed up the algorithm by, in some cases, an order of magnitude by limiting the heights to be searched. This gain is derived from simulations and is achieved whilst at the same time not neglecting large search volumes, continuing to allow a change of audio sources to be detected.

Index Terms— Microphones, Acoustic measurements, Optimization methods, Sampling methods

1. INTRODUCTION

Acoustic source localisation has been studied extensively in the literature [1–3]. Systems make use of an array of microphones to sample audio data and commonly use time difference of arrival (TDOA) techniques to estimate the angle of arrival of a sound wave relative to a pair of microphones. These angles can then be used to triangulate the location of an acoustic source [4]. The steered response power (SRP) is a slower method which also has potential for use in multi-speaker detection. This paper relates to previous work by building on a successful technique which uses the SRP to find an audio source quickly by reducing the number of calculations needed to localise a source. The work presented reduces this number further, making the algorithm useful even in relatively low signal to noise ratio (SNR) environments.

The SRP is a useful measure of the acoustic power originating from a particular location in space within a room. It has been shown to be relatively robust to reverberation. The generalised cross correlation with phase transform (GCC-PHAT) [5] from a set of microphones is used by the SRP algorithm to build up a 3-dimensional (3D) map of this power. Since the volume of a room is very large compared to the spatial resolution generally required by source tracking applications and because of the slow nature of the algorithm, the calculation of the SRP across an entire room is computationally expensive. The output is also a 3D array, which makes it costly, although not intractable, to search through. There are various methods [6–8] for finding global maxima of the array, however SRP based audio localisation also has the potential to locate and track multiple speakers more easily than the traditional maximal generalised cross correlation (GCC) TDOA methods [9].

Existing work reduces the time taken to find a maximum within an area by sampling from the search space randomly and then recursively shrinking the search space using the best subset of the results, a technique called stochastic region contraction (SRC) [7]. Rather than assuming the search volume is the whole room, the SRC algorithm [10] assumes that the height of the search volume is restricted to being one metre high and is also offset from the ground [7]. The contribution of this work is a method to extend the implicit assumptions of head height made when using the SRC by assuming that prior information of the expected head height at some positions is available. For example, this information can be estimated from a camera system using Viola-Jones face detection [11, 12]. The contribution in 3 then interpolates and extrapolates to estimate head height across the 2-dimensional (2D) search area, $\mathbb{A}$, and from that, a sampling distribution over height is formed across the room. This allows the number of functional evaluations (FEs) required to find a maximum to be reduced. Because interpolation is used to estimate head height across an area, people missed in the visual domain due to occlusion are still quickly locatable in the audio domain.

This paper describes the SRP, which is the functional of the SRC algorithm, and then goes on to describe the interpolation and probability density function (PDF) used in the proposed height estimation (HE) SRC algorithm. This algorithm is then tested on a recorded data set which had the room set up, for comparison, to be similar to the conditions described in [7]. The paper also proposes a novel approach to multi-source audio localisation. By sampling across every 2D point

The authors would like to give thanks for Scholarships from the James Clerk Maxwell Foundation and from the Maxwell Advanced Technology Fund at the University of Edinburgh.

within a room at a height drawn from this distribution, a 2D SRP map can be made of the search area at relatively low computational cost. This may prove itself to be useful for algorithms to find multiple maxima, corresponding to multiple audio sources, for robust multi-speaker localisation. By increasing the number of samples at each height and averaging, this tends towards the marginalisation of the SRP over height.

2. STOCHASTIC REGION CONTRACTION

A popular method of audio source tracking is extracting and triangulating TDOA values from the maxima of the GCC-PHAT of signals from pairs of microphones in the frequency domain, given by Equation (1)

$$\hat{R}_{x_m x_n}(\tau) = \int_{-\infty}^{\infty} \frac{\hat{G}_{x_m x_n}(f)}{|\hat{G}_{x_m x_n}(f)|} e^{2\pi f \imath \tau} \, df \qquad (1)$$

which is an inverse Fourier transform where $\hat{G}_{x_m x_n}$ is the product of the signals x_m and x_n in the frequency domain.

The SRP makes use of the GCC-PHAT to build an energy map for each point $(x,\, y,\, z)$ in a search area $\mathbb{A}$ using Equation (2)

$$S(x, y, z) = \sum_{n=1}^{M} \sum_{m=n+1}^{M} \hat{R}_{x_n\, x_m}[\tau_{n\, m}(x, y, z)] \qquad (2)$$

in a system with M microphones. This is the sum over all pairs $(m,\, n)$ of microphones of the corresponding value of the GCC-PHAT for the TDOA τ. The TDOA is defined by Equation (3)

$$\tau_{n\, m}(\mathbf{p}) = (|\mathbf{m} - \mathbf{p}| - |\mathbf{n} - \mathbf{p}|) \,/\, c \qquad (3)$$

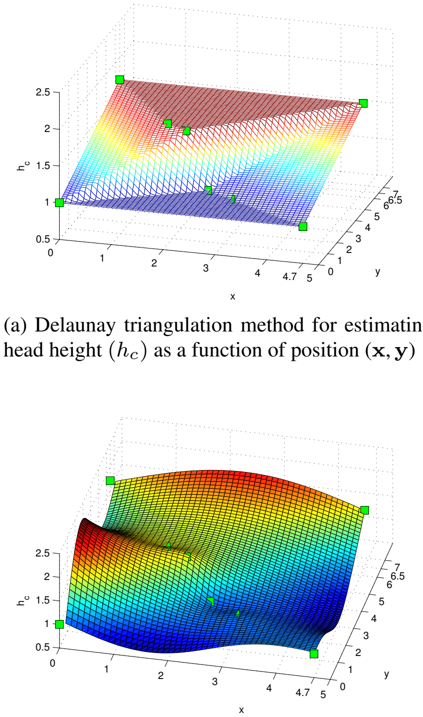
where $\mathbf{p}$ is the vector $(x,\, y,\, z)$ of the point under investigation, c is the speed of sound, and $\mathbf{m}$ and $\mathbf{n}$ are the positions of microphones m and n respectively.

SRC takes samples of the SRP from across the search space and attempts to contract it by using the area given by a set of the highest valued samples [7]. Because these will generally be centred around a peak, caused by a sound source, the search area should quickly shrink. By repeating this, the search space will become an area sufficiently small enough to be considered the point which is the maximum of the SRP function and therefore the source of the sound.

3. INTERPOLATION

To choose head height, existing knowledge of the current positions and heights of people in a room can be used. In an audio-visual (AV) system, this is easy to initialise as video data can be used to make an initial estimation of the heights which should be searched in the audio domain. In addition, existing audio domain search techniques such as the full SRC algorithm can be used to make the first head height estimation. After they have been found initially, the tracked locations of people, both speakers and non-speakers, from both audio and visual sources will allow a good estimate of the height to be used across the room. From a sparse set of people, the head height to be used at every x-y co-ordinate in the SRP map needs to be defined. This means that an assumption about the outer elements of the set and how they relate to the height at the edge of the search area must be made. This work uses the speaker closest to a corner to specify the height at that corner.

When doing interpolation, there is a trade-off between the smoothness of the curve produced and the size of ripples produced. The interpolation should not contain severe ripples as they would lead to large errors in the head height estimation across the room. Ideally, it should be monotonic and one way to achieve this is to use Delaunay triangulation [13] on the set of speakers, which creates a surface which can be evaluated at any 2D point. Figure 1 compares Delaunay triangulation based interpolation to a plate-splines method [14], where the room dimensions are along the x and y axes and the interpolated heights h_c form the set $\mathbf{H}$ across the area of the room. These show that the Delaunay method solves the problem of large ripples, although it leads to a less smooth interpolation. In order to extrapolate correctly, room corners must be pre-allocated nodes. There are several options for choos-



(a) Delaunay triangulation method for estimating head height (h_c) as a function of position $(\mathbf{x}, \mathbf{y})$

(b) Plate-splines method for estimating head height (h_c) as a function of position $(\mathbf{x}, \mathbf{y})$

Fig. 1. Interpolation Method Comparison ing the height h_{c_j} at each of these j nodes (in a rectangular room, $j = 4$), such as choosing the height to be the same as the height of the nearest speaker, as shown in Equation (4a), where z_i is the height component of r_i, the position of known node i and r_{c_j} is the position of corner j. An alternative is to use Equation (4b), the expected height of a speaker from all known node heights z_i. If it is assumed that there are a limited number of speakers then finding the nearest node to a corner poses no computational problems.

$$h_{c_j} = \arg\min_{z_i}[r_{c_j} - r_i] \tag{4a}$$

$$h_{c_j} = \mathbb{E}\left[z_i\right] \tag{4b}$$

Because the head height, $\mathbf{H}$, is only an estimate, its accuracy varies across the room. To compensate, the head height to be used in the SRC algorithm is drawn from a PDF which ensures that most of the time, samples are taken around head height without being overly restrictive and a small amount of time from less likely areas, so as not to entirely neglect large portions of the search space. The interpolated head height is taken as the mean of a Gaussian distribution whose variance changes depending on its proximity to a known source. This allows the search to concentrate on areas likely to contain people whilst at the same time, not neglecting to check for possible outliers. The height h_{sub} to use at each time step for every 2D point $\mathbf{p_2} = (x_{p_2},\, y_{p_2})$ is then drawn from (5) where $\mathbb{T}$ is the set of known speaker locations.

$$\begin{aligned} \varphi\left(z \mid \mathbf{p_2}\right) &= \alpha_0 \mathcal{N}\left(\mu_h,\, \sigma_h^2\right) + (1-\alpha_0)\,\mathcal{U}\left(0,\, h_r\right) \\ \mu_h &= \mathbf{H}[\mathbf{p_2}] \\ \sigma_h^2 &= \hat{q}(\mathbf{p_2},\, \mathbb{T}) \end{aligned} \tag{5}$$

which mixes the Gaussian with a Uniform distribution across h_r, the entire height of the room.

This can be repeated n times to create an array where $h[n] = h_{sub}$ each time. The resulting SRP value for the point $\mathbf{p_2}$ can either be the maximum value found as in Equation (6a) or the expectation (Equation (6b))

$$SRP_{\mathbf{p_2}} = \max_z [S(x_{p_2}, y_{p_2}, h[n])] \tag{6a}$$

$$SRP_{\mathbf{p_2}} = \mathbb{E}\left[S(x_{p_2}, y_{p_2}, h[n])\right] \tag{6b}$$

of the values, in which case as n increases, $SRP_{\mathbf{p_2}}$ tends towards the marginalisation of the SRP over z, the room height.

Around each person, we can be relatively confident of their height. Further away from them, the decreasing confidence is modelled by increasing the variance of the sampling PDF. The variance at a distance l metres from a speaker is chosen to be modelled by a sigmoid function, q, such as Equation (7a), which is a scaled error function, or Equation (7b).

$$\mathrm{q}(l) = \alpha_1 \operatorname{erf}\left(\alpha_2 l\right) \tag{7a}$$

$$\mathrm{q}(l) = \alpha_1 (1 - e^{-l/\alpha_2}) \tag{7b}$$

These are both 0 at the origin and asymptotically approach constants as their arguments tend towards infinity.

These are combined to form a global variance in Equation (8).

$$\begin{aligned} \mathbb{L}_{\mathbf{p},\mathbb{T}} &= \{l : (\exists \mathbf{q} \in \mathbb{T})(l = |\mathbf{p} - \mathbf{q}|)\} \\ \hat{q}(\mathbf{p_2},\, \mathbb{T}) &= \min_{l \in \mathbb{L}_{\mathbf{p},\mathbb{T}}} q(l) \end{aligned} \tag{8}$$

At any point $\mathbf{p}$ in space, the appropriate variance $\hat{q}$ to use will be the sigmoid function q of the minimum of the set of all 2D Euclidian distances $\overline{\mathbf{pq}}$ to known sources, where an element of $\mathbb{T}$ is denoted as $\mathbf{q}$. The minimum is chosen to ensure that the change in variance remains smooth even for overlapping sigmoids from multiple sources.

4. ALGORITHM

The algorithm for finding the global maximum using the estimated head height is given in Algorithm 1, where DT is the Delaunay Triangulation operation.

```
Initial search for a speech source
while $running$ do
    $\hat{\mathbb{T}} = \mathbb{T}$
    for all room corners do                     ▷ Add room corners to $\hat{\mathbb{T}}$
        $\mathbf{n} \leftarrow (x_{\text{corner}},\, y_{\text{corner}},\, z_{\text{nearest member of } \mathbb{T}})$
        $\hat{\mathbb{T}} \leftarrow \hat{\mathbb{T}} \cup \{\mathbf{n}\}$
    end for
    $\hat{\mathbb{H}} \leftarrow \mathrm{DT}(\hat{\mathbb{T}})$              ▷ Delaunay Triangulation of the set
    for all $\mathbf{p_2} = (x_{p_2},\, y_{p_2}) \in \mathbb{A}$ do    ▷ Whole search area
        $\hat{\mathbb{H}}_0 \leftarrow h_{sub} \sim \varphi\left(z \mid \mathbf{p_2}\right)$          ▷ Choose a height
    end for
    Perform SRC with heights from $\hat{\mathbb{H}}_0$
    $\mathbb{T} = \mathbb{T} \cup \{$new speaker positions$\}$
end while
```

Algorithm 1: HE-SRC Algorithm

5. EXPERIMENTAL RESULTS

The algorithms were run in the environment shown in Figure 2, where the red circles represent each of the 12 microphones (placed along the edges of the room, similar to the panels used in [15]) and the green squares represent the speaker positions. This was a (4.7x6.5)m room, as described in [15] in order to make a direct comparison. A minute of data was recorded for each speaker at 96,000kHz, which gave each around 300 audio windows based on a window size of 160ms. Speakers did not talk at the same time and the two speakers furthest away from the array were at the lower height of 1m, rather than 1.6m, in order to show that this doesn't affect the algorithm. The variant of SRC used was SRC-I, which fixes J_0 - the number of points to be evaluated at the first iteration

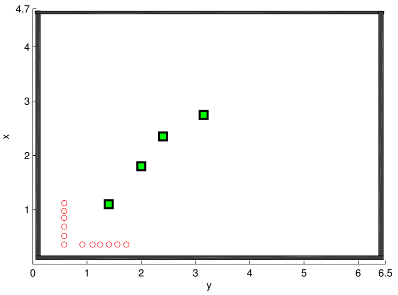


Fig. 2. Room Layout

- to a constant J and then calculates J_i FEs at each iteration of the algorithm, which is decided dynamically [7]. In this variant, a number N of the highest valued samples are used to contract the search region [7]. α_0 was chosen to be 0.95 in order to concentrate the search within head height. Lower values weight the distribution to uniformly draw from across the height of the room, making the search similar to the original SRC algorithm, but with fewer assumptions and therefore slower searches. α_1 was chosen to be 0.5, allowing most of the Gaussian distribution to concentrate on an area 1m tall, similar to the 1m tall Uniform distribution used for height in the original SRC algorithm. Finally, α_2 was generated by choosing the radius l, at which the sigmoid function should be 99% of the way towards α_1, to be 1m, which assumes people have some personal space whilst talking.

Data was evaluated using an Average Location Error (ALE) - the mean of the Euclidian distances of each set of results to their corresponding ground truths. Because the search space was reduced by the height estimation, the number of samples J_i at each stage was lowered to improve overall search times, trading off against accuracy. In the first instance, HE-I, only 350 samples were taken at the first iteration with only the top $N = 30$ used for region contraction. Accuracy decreased as the sound source was further away from the microphone array, implying a lower SNR as in [7], but this may be acceptable in a system whose tracker accounts for noisy state observations and exploiting this may warrant further investigation. For HE-II, J_0 was set to 1000 and N to 60, which brought the accuracy across all sources up whilst keeping the number of FEs low. In HE-III, J_0 was set to 3000 and N to 60, the value as used in [7]. Table 1 shows the results of first (SRC-I) variation of the SRC algorithm from [7] on the data set and compares these configurations with the HE variants. It shows the average number of FEs used within an audio frame and the ALE, where Source 1 is the closest to the microphone array and Source 4 is the furthest.

The results show that with prior information about head height within a room, the SRC can be sped up whilst maintaining accuracy. Because in HE-III the parameters are similar to the SRC-I parameters, the algorithms are expected to perform similarly when there is no known audio source. In this case, the mean of the Gaussian is set to the same offset as that used in the algorithm and the variance is again set to 0.5.

Algorithm	**Source 1**		**Source 2**		**Source 3**		**Source 4**	
	ALE (m)	**# FEs**	**ALE (m)**	**# FEs**	**ALE (m)**	**# FEs**	**ALE (m)**	**# FEs**
SRC-I	0.26	61,1001	0.31	61,1001	0.45	61,001	0.6	61,001
HE-I	0.32	17,156	0.35	21,939	0.44	31,811	0.58	35,053
HE-II	0.12	34,022	0.22	35,136	0.26	41,228	0.5	39,402
HE-III	0.11	40,721	0.15	40,736	0.23	42,900	0.34	44,111

Table 1. Comparison of SRC Methods

Algorithm	**Source 1**		**Source 2**		**Source 3**		**Source 4**	
	ALE (m)	**# FEs**	**ALE (m)**	**# FEs**	**ALE (m)**	**# FEs**	**ALE (m)**	**# FEs**
HE-I	0.45	20,115	0.49	23,849	0.51	36,253	0.6	37,962
HE-II	0.23	35,117	0.24	36,140	0.41	47,281	0.47	48,294
HE-III	0.22	43,783	0.24	43,548	0.32	55,352	0.46	56,667

Table 2. FEs required to find a source with no prior

Table 2 shows the average number of FEs required to find a source using the algorithm without prior information. The results indicate a reduced performance with HE-III, but still within the tractable range of tens of thousands of FEs and close to the performance of SRC-I, as expected. For lower values of J_0 and N, results are improved. In particular, HE-II provides good accuracy and good performance, with or without prior information, so much so that it is suitable as an audio estimator for the initial height information in this situation.

6. CONCLUSIONS

This work contributes a method of speeding up and increasing the accuracy of the SRC algorithm by estimating the height at which to search from prior information, obtainable via standard methods and information from a previous iteration of the algorithm. The key to this technique is to estimate an average head height across an area by interpolating and extrapolating heights of known speakers and forming a probability distribution of head height using this data. This allows a single audio source to be localised quickly whilst still searching across the room to find new source, for example when there is a speaker change. Further work will investigate using the height estimated SRP to locate multiple maxima simultaneously.

7. REFERENCES

[1] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, oct. 2009, pp. 2033 –2038.

[2] Weiping Cai, Shikui Wang, and Zhenyang Wu, "Accelerated steered response power method for sound source

localization using orthogonal linear array," *Applied Acoustics*, vol. 71, no. 2, pp. 134 – 139, 2010.

[3] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1409 –1415, may 2012.

[4] Matthias Wölfel and John. McDonough, *Distant Speech Recognition*, Wiley, 2009.

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

[6] J.P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2510 –2526, nov. 2007.

[7] Hoang Do, H.F. Silverman, and Ying Yu, "A real-time SRP-PHAT source location implementation using Stochastic Region Contraction (SRC) on a large-aperture microphone array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, april 2007, vol. 1, pp. I–121 –I–124.

[8] Hoang Do and H.F. Silverman, "Stochastic particle filtering: A fast srp-phat single source localization algorithm," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, oct. 2009, pp. 213 –216.

[9] Fotios Talantzis, Aristodemos Pnevmatikakis, and Anthony G Constantinides, *Audio-Visual Person Tracking: A Practical Approach*, Imperial College Press, 2012.

[10] M.F. Berger and H.F. Silverman, "Microphone array optimization by stochastic region contraction," *Signal Processing, IEEE Transactions on*, vol. 39, no. 11, pp. 2377–2386, 1991.

[11] Paul Viola and Michael Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[12] T. Gehrig, K. Nickel, H.K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, oct. 2005, pp. 118 – 121.

[13] D. T. Lee and B. J. Schachter, "Two algorithms for constructing a Delaunay triangulation," *International Journal of Parallel Programming*, vol. 9, pp. 219–242, 1980, 10.1007/BF00977785.

[14] John D'Errico, "inpaint_nans," MATLAB Central File Exchange, February 2012.

[15] H.F. Silverman, Ying Yu, J.M. Sachar, and II Patterson, W.R., "Performance of real-time source-location estimators for a large-aperture microphone array," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 593 – 606, july 2005.