

MULTIMODAL LOCALIZATION OF A FLYING BAT

Kaushik Ghose ^a, Dmitry Zotkin^b, Ramani Duraiswami ^b and Cynthia F. Moss^a

a: Auditory Neuroethology Laboratory,

b: Perceptual Interfaces and Reality Laboratory, UMIACS

University of Maryland College Park, MD 20742

{kghose@wam,dz@cs,ramani@umiacs,cmoss@psyc}.umd.edu

ABSTRACT

In this paper we present a new multimodal system that combines stereoscopic and audio-based source localization to perform behavioral studies on a flying bat. Also presented are novel algorithms for audio source localization. The bat was allowed to fly in an anechoic flight room and monitored by two high speed video cameras. The vocalizations of the bat were simultaneously recorded from six microphones. The data was then processed offline to localize the source and reconstruct the trajectory of the bat. We compare the performance of the localization algorithm with the position data obtained from stereoscopic pictures of the bat. The results confirm that the stereoscopic analysis and the audio localization are in good agreement. This system opens up new possibilities for performing multimodal research, and developing more tightly integrated algorithms.

1. INTRODUCTION

Combining audio and video based tracking is the stated goal of many systems. In this paper we present a laboratory system that combines audio and video source localization for studying the behavior of echolocating bats. Our goal is to mutually validate the two modalities of source detection, and to build a general system that tracks fast moving objects in a room. An interesting aspect of the present application, is the use of mutual validation among the localization techniques, that enables Another goal of the paper is to present some novel algorithms for acoustic source localization. In §2 we introduce the problem under consideration, in §3 we introduce the audio source localization algorithms used. In §4 the video localization algorithms (stereo) are introduced. The paper concludes with experimental results and discussion in §5.

2. PROBLEM: BAT BEHAVIORAL STUDY

Echolocating bats actively probe the environment by producing ultrasonic vocal signals (short chirps) consisting of a constant frequency (CF) signal and/or a frequency modulated (FM) signal. These chirps reflect from objects in the path of the sound beam and the bat uses information contained in returning echoes to determine the direction, distance, size and possibly shape of sonar targets. Thus, in echolocating bats, active sonar replaces vision as a modality for navigation and hunting [1]. The bat biosonar, due to its sophistication, serves as an excellent model for studying auditory localization in animals [2]. Studies on free flying bats have to be carried out under controlled conditions (total darkness) so that the possibility that the bat may be using visual cues is eliminated. It is important to record both the vocalizations as well as the flight path of the bat in order to gain a thorough understanding of the

bat's behavior. Presently, under laboratory conditions, it is possible to do this in a limited way by using two high speed infra red sensitive cameras to record the bat's flight and then reconstructing the 3 dimensional flight path using stereoscopic techniques. The recorded vocalizations of the bat are then matched in time with the flight path reconstruction. The disadvantage to this technique is that, **a)** This is only possible under very controlled conditions, i.e. only in a large flight room with carefully positioned cameras, **b)** there is a fairly restricted volume within which the path may be reconstructed accurately, and the bat often spends a great deal of time outside this volume, and a lot of interesting behavior can not be studied quantitatively. For instance, during hunting there are typically three main stages the bat goes through: search, approach and capture. This behavior is studied in the lab by training bats to catch prey suspended in view of the cameras. The search and early approach phases often take place outside the camera view, so the flight behavior during this phase is not quantifiable.

Using an array of microphones and source localization techniques it is possible to locate the bat whenever it makes a vocalization. This method of locating the bat is feasible as long as the bat's vocalization is loud enough to be picked up by at least 4 microphones in the array. This approach enables us to improve upon current methods of studying bat behavior by enabling the bat to be tracked over a greater extent in space (and also time). This enables us to locate the bat even during the search and early approach phases, potentially revealing interesting details of flight planning during this stage.

Experimental technique The bat used in this study, *Eptesicus fuscus*, emits ultrasonic chirps consisting solely of downward sweeping FM sounds. The signal bandwidth extends from 50 kHz to 20 kHz for the fundamental component. The duration of the signals range from 20 ms down to 2 ms. The bat was trained to fly in a large (5m x 5m x 2.5 m) anechoic room and capture a mealworm suspended from the ceiling by a microfilament. The bat's flight was recorded using two Kodak MotionCorderTM digital cameras running at 240 Hz. Vocalisations of the bat were recorded from six microphones (Knowles FG3329) arranged in an "L" shaped array. Sounds were digitized at 140 kHz/channel using an IoTech WavebookTM. The video and audio data were synchronized by running the acquisition off a common trigger. A schematic of the room is shown in Figure 1.

3. STEREOSCOPIC LOCALIZATION

The projective camera model [4] indicates how a 3D world point $\mathbf{X} = (x, y, z)^t$ is projected to the point $\mathbf{U} = (u, v)$ in the camera

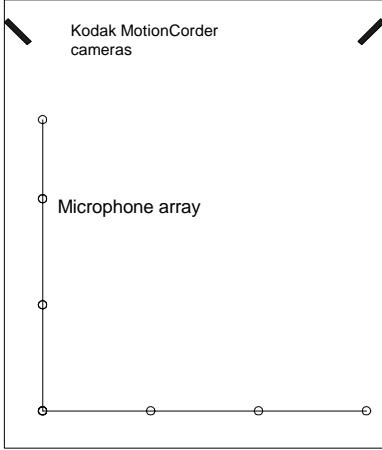


Fig. 1. Schematic of flight room experimental set up.

plane

$$[\mathbf{K}|\mathbf{0}] \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{U} \\ 1 \end{bmatrix} \quad (1)$$

Here \mathbf{K} is the camera intrinsic calibration matrix (focal length, skew, pixel stretch ratios), while \mathbf{R} and \mathbf{t} are the camera extrinsic calibration parameters, indicating the rotation and translation necessary to reach the frame at the camera center from the world origin. Determining these parameters constitutes the camera calibration step. We calibrate the cameras using a calibration rig that provides 25 unique points in a region that occupies a $[2\text{m}]^3$ volume. The world coordinates of these points are known to an accuracy of 5 mm. Using these points, the Peak Motus system uses a Direct Linear Transformation algorithm [4] to obtain the camera calibration.

Using a pair of well separated cameras, we image a portion of the room that includes the bat flight path. For points that are observable from both camera locations we obtain the three dimensional coordinates via stereo. To perform accurate reconstruction using the video system, we record video events of bats flying and hand segment the video to locate the bats in the images of the two cameras. Care is taken to ensure that the corresponding points are from parts of the object that deform rigidly (i.e. not on the batwings.)

Stereo is known to be prone to errors, especially at the wide baselines that are used in the present case. Further, only a small portion of the space is captured by both cameras.

4. AUDIO ALGORITHMS

It would thus be useful to compare the stereo data with other means. In the present case time delays at the microphone array used to record the bat vocalization's directivity can also be used to estimate the source position at the instants the bat is emitting sound.

Determining the source coordinates from measured time differences is an almost classical problem arising in many different fields of signal processing. We have N microphones located at points $\mathbf{m}_i = (x_i, y_i, z_i)$, and a source at $\mathbf{s} = (x_s, y_s, z_s)$. The

speed of sound is denoted c , and distances between the microphones and the source is indicated as χ_i , with

$$\chi_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2 + (z_i - z_s)^2}. \quad (2)$$

The measured time delays between microphones i and j each provide a linear relationship of the form

$$\chi_i - \chi_j = ct_{ij}. \quad (3)$$

In general for N microphones there are $C(N, 2)$ measurements of which $N - 1$ are independent. We obtain the time delays using a robust algorithm that uses the noise estimate in the absence of the signal as a weight [3].

Exact solution For our “L”-shaped microphone array configuration we can employ a novel exact solution [8]. We consider one arm of the array, and set our origin at the microphone common to the two arms. We take two additional microphones along the arm, with spherical coordinates $(R, 0, 0), (2R, 0, 0)$. For a given source at (r, θ, ϕ) , we denote the distance between the source and microphone i as χ_i . Then $\chi_1 = r$, and

$$\chi_2^2 = r^2 + R^2 - 2rR \cos \theta, \quad \chi_3^2 = r^2 + 4R^2 - 4rR \cos \theta. \quad (4)$$

Three microphones give us two unique time delays, thus this configuration cannot be used to determine ϕ . However one can determine r and θ , and the determined source location lies on a circle as ϕ varies between 0 and 2π . To determine the χ_i we can use the two unique timedelays, taken as t_{12} and t_{23} , and get two equations of the form (3). In addition we can write the following identity

$$2\chi_2^2 - \chi_1^2 - \chi_3^2 = -2R^2 \quad (5)$$

This nonlinear constraining can be made linear by using the time delay expressions and written as

$$-ct_{12}(\chi_2 + \chi_1) + ct_{23}(\chi_2 + \chi_3) = -2R^2$$

The resulting system can be solved for the χ_i as

$$\begin{bmatrix} \chi_1 \\ \chi_2 \\ \chi_3 \end{bmatrix} = \frac{1}{t_{13}} \begin{bmatrix} -2t_{23} + t_{12} \\ -\frac{t_{12}^2}{2}c - \frac{t_{23}^2}{2}c + \frac{R^2}{c} \\ -\frac{t_{12}^2}{2}c - \frac{2t_{12} - t_{23}}{2}ct_{23} + \frac{R^2}{c} \end{bmatrix}$$

We can get the range and the coordinate x_s as

$$\langle r \rangle = \frac{\chi_1}{2} + \frac{1}{2} \sqrt{\frac{2\chi_2^2 - \chi_3^2 + 2R^2}{2}}, \quad (6)$$

$$x_s = \langle r \cos \Theta \rangle = \frac{3\chi_1^2 - \chi_3^2 + 12R^2}{12R} \quad (7)$$

We can now use the common microphone and 2 microphones along the other arm of the L to get the y coordinate, and consequently, the full source location.

Source localization algorithm The above exact solution requires accurate time delays and sound speeds for reasonable performance. In addition it does not make use of measurements from all M microphones. For more robust performance in the presence of noise and outliers we use a second novel algorithm [7]. This is based on the observation that the source location estimation can be decomposed into two independent sub-problems. The first sub-problem

involves the measured time differences (3) which involve potential errors due to multipath and reverberation, and due to errors in the sound speed value. This sub-system has rank $M - 1$. We make the definition

$$\mathbf{d} = [\chi_2 - \chi_1, \dots, \chi_M - \chi_1]^t. \quad (8)$$

so that the independent set that must be estimated from the noisy measurement can taken to be \mathbf{d} .

We can estimate \mathbf{d} by solving the rank-deficient problem by imposing hard constraints that impose $R_{max} > \chi_i \geq 0$, and also bound time delays, and incorporate knowledge of the expected imprecision in the measurements to throw out outliers. These constraints have the form

$$\chi_i - \chi_j > c_{\min}(t_{ij} - \epsilon), \quad \chi_i - \chi_j < c_{\max}(t_{ij} + \epsilon), \quad (9)$$

for $t_{ij} > 0$, and with similar equations for $t_{ij} < 0$. This set of equations (3) and (9) is solved using a constrained L_1 optimization algorithm, termed “CL1” [6]. Solving the above equations with CL1 yields a solution with the value to the closest microphone as zero, i.e. we arrive at a constrained L_1 norm estimator for \mathbf{d} in Equation (8) above using all the measurements, but excluding those outliers that violate constraints.

Knowing \mathbf{d} , in the second stage of our solution, we estimate χ_1 and the coordinates using the procedure of Smith and Abel [5], except that we begin with an improved estimate of \mathbf{d} . We make a few definitions for the Smith Abel solution. Let R_i be the distance between microphone i and microphone 1, i.e. $R_i = |\mathbf{m}_i - \mathbf{m}_1|$, and let

$$\mathbf{S} = \begin{bmatrix} \mathbf{m}_2 - \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_M - \mathbf{m}_1 \end{bmatrix}, \quad \delta = \begin{bmatrix} R_2^2 - d_1^2 \\ \vdots \\ R_M^2 - d_{M-1}^2 \end{bmatrix}$$

The Smith Abel solution for the unknowns (χ_1 and \mathbf{x}_s) is

$$\mathbf{S}_W^* \equiv (\mathbf{S}^t \mathbf{W} \mathbf{S})^{-1} \mathbf{S}^t \mathbf{W}, \quad \mathbf{P}_s = \mathbf{S} \mathbf{S}_W^*, \quad \mathbf{P}_S^\perp = \mathbf{I} - \mathbf{P}_s \quad (10)$$

$$\chi_1 = \frac{\mathbf{d}^t (\mathbf{P}_S^\perp (\mathbf{W} (\mathbf{P}_S^\perp \delta)))}{2\mathbf{d}^t (\mathbf{P}_S^\perp (\mathbf{W} (\mathbf{P}_S^\perp \mathbf{d}))), \quad \mathbf{x}_s = \frac{1}{2} \mathbf{S}_W^* (\delta - 2\chi_1 \mathbf{d}), \quad (11)$$

where \mathbf{W} is a weighting vector, which is assumed as identity in the present computations. Finally, after obtaining the Smith-Abel estimate, we perform local function minimization using the Euclidean distance between the vector \mathbf{d} as provided by CL1 and as obtained from the computed source coordinate position as the objective function, using a standard routine fminsearch from MATLAB.

5. RESULTS

We present results from two trials recorded of a bat moving towards a mealworm prey. There is also an inedible distractor located in proximity to the edible target. The bat flies in from the right towards the target located at the left in these figures (which show a plan view of the room). The behavior of the bat is as follows. In the beginning of the trial the bat is in general flight (search mode), and only emits infrequent vocalizations. As the bat acquires the target it begins emitting more frequent vocalizations

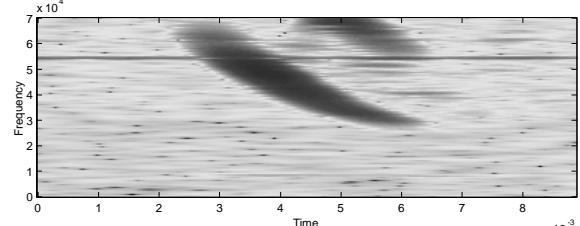


Fig. 2. Spectrogram and waveform of a bat vocalization.

(approach). After the capture, indicated by a joining of the estimates of the bat’s track and the target track, the bat is silent for a while, and then begins to emit search mode clicks again. The density of the audio estimations provides thus both behavioral data and localization data. A spectrogram of a typical bat vocalization is shown in Figure 2, and is the signal that is used in the localization. The first set of figures shown below provide the estimates from the exact audio solution, compared to the stereoscopic software output. As can be seen, the exact solution results track the video data quite well. There are a few outliers (which could be easily removed by a posteriori estimates of the delay at other microphones) which are included in the picture, with the purpose of showing that the exact solution can fail when there is error. In figure 4 we show the performance of the CL1 algorithm for the same data. Also shown are error estimates (the distance between the vector \mathbf{d} estimated by CL1 and that from the estimated source position). Note that this distance grossly overestimates the error, and should be divided by the number of microphones.

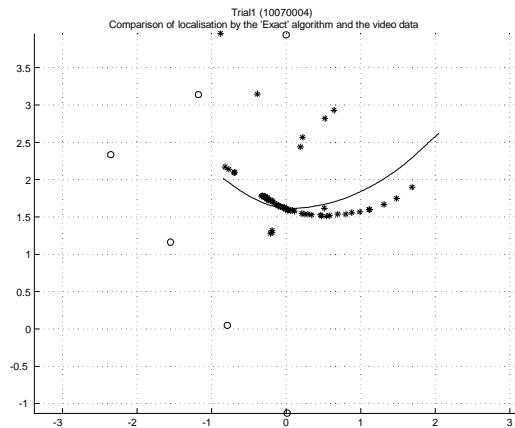


Fig. 3. Exact solution results (*) and stereoscopic data (solid line) for trial 1.

A second trial is shown in Figures 5 and 6. In this trial both the

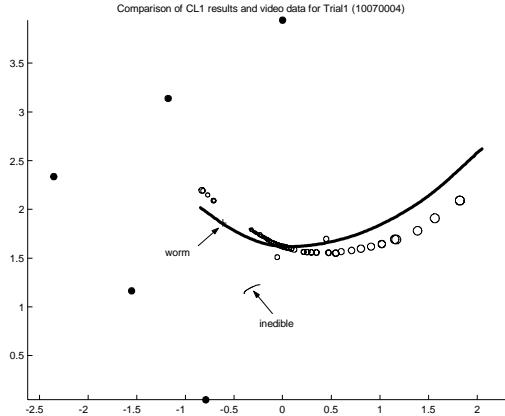


Fig. 4. Comparison of CL1 results and video data. The circles show the positional error estimates of the CL1 results.

prey and the distractor are moving, also from right to left. The bat is able to come near the correct target, but misses it. Both the exact solution and the CL1 algorithm do reasonably well in capturing the bat's motion. However, there again are many more outliers in the exact solution. These could be easily be eliminated by temporal filtering or a posteriori verification of the delay data, but have not been in these graphs.

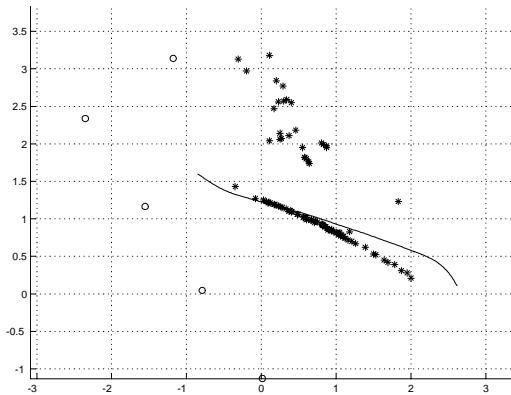


Fig. 5. Exact solution results for the 2nd trial.

6. CONCLUSIONS

Based on these preliminary studies we can make the following conclusions.

- Audio and video provide complementary modalities to investigate the behavior of a bat in flight.
- The audio algorithms are suitable for real-time tracking over a much wider field of view.
- Despite our expectations to the contrary, for this configuration, the exact solution gives as good results as the more complex CL1 algorithm. In addition the exact solution only uses triples of microphones, which restricts the range of the lags of the cross-correlation, resulting in a faster and more robust estimate of time delays.

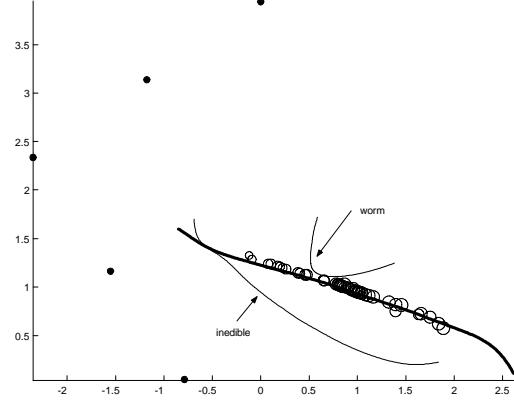


Fig. 6. As in Fig. 4, but now with a moving target and distractor. The bat approaches the correct target, but misses it.

- The differences in the video and audio data are larger when the track is closer to the right boundary. This region is known to be poorly estimated by the stereoscopic analysis because it lies in the more distorted portions of the picture, and is farther from the calibrated region of space.

7. REFERENCES

- [1] D.R. Griffin (1958) "Listening in the Dark", *Yale U.P.*
- [2] A. Popper & R. Fay ed. (1995) "Hearing by bats", Springer Handbook of Auditory Research, Vol. 5, Springer..
- [3] D.Feitelson & A.Weil (1996). "A robust method for speech signal time-delay estimation in reverberant rooms", *Proc. ICASSP-96, Atlanta, GA.*
- [4] R. Hartley & A. Zisserman, (2000) Multiple View Geometry in Computer Vision, *Cambridge*.
- [5] J.Smith & J.Abel, (1987) "Closed-form least-squares source location estimation from range-difference measurements", *IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-35(12)*, 1661-1669.
- [6] I. Barrodale and F.D.K. Roberts, (1973) "An improved algorithm for discrete l_1 linear approximation," *SIAM J. Numer. Anal.*, **10**, 839-848.
- [7] D. Zotkin, R. Duraiswami, L.S. Davis & I. Haritaoglu.(2000) An audio-video front-end for multimedia applications, *Proc. IEEE SMC 2000, Nashville, TN.*
- [8] R. Duraiswami, D. Zotkin, & L.S. Davis (1999) Exact solutions for the problem of source location from measured time differences of arrival. *J. Acoust. Soc. Am.*, v. 106, p. 2277.