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

We present a new multimodal system that combines stereoscopic and audio-based source localization to track the position of a flying bat. Also presented are novel algorithms for audio source localization. The bat was allowed to fly in an anechoic 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 steroscopic 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 free flying echolocating bats. Our goal is to mutually validate the two modalities of source localization, 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. 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 and §4 respectively we introduce the video and the audio source localization algorithms used. The paper concludes with experimental results in §5 and conclusions in §6.

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 [1]. Thus, in echolocating bats, active sonar can replace vision as a modality for navigation and hunting [2]. The bat biosonar serves as an excellent model for studying auditory localization in animals [3]. Studies on free flying bats have to be carried out under controlled conditions (a dark room lit with IR lamps) 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 region, and a lot of interesting behavior can not be studied quantitatively. For instance, insect capture behavior can be characterized by three phases in a sequence: 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 specify the bat's position in space whenever it makes a vocalization. This method of locating the bat is feasible as long as the bat's vocalization is strong 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 behavioral planning well before target interception.

Experimental technique The bat used in this study, *Eptesicus fuscus*, emits ultrasonic chirps consisting of downward sweeping FM sounds. The signal bandwidth extends from 60 kHz to 25 kHz for the fundamental component. The duration of the signals range from 20 ms down to 0.5 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 MotionCorder^{TM} digital cameras running at 240 Hz. Vocalizations 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 Wavebook^{TM}. The video and audio data were synchronized by running the acquisition off a common trigger. A schematic of the setup is shown in Figure 1.

3. STEREOSCOPIC LOCALIZATION

We are given a pair of widely-spaced cameras that can view the space under consideration, and a calibration object of known shape and size. For determining the 3-D position of points from approximate correspondences, the simplest algorithm to use was a classical one described in Slama [10], that is also extensively used in

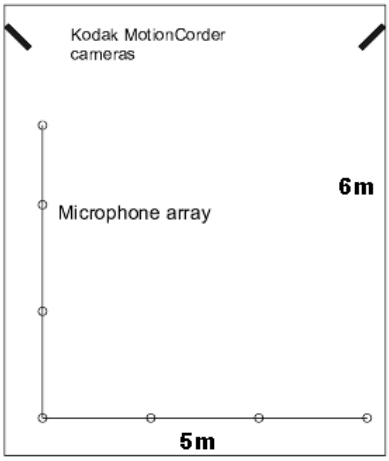

Fig. 1. Schematic of flight room experimental set up.

the gait-analysis and motion capture communities. This algorithm is also discussed in Chapter 11.2 of [5].

We calibrate the cameras using a calibration frame that provides 25 unique points in a region that occupies approximately a [2m]^3 volume. The world coordinates of these points are known to an accuracy of 5 mm. Using the known coordinates of the calibration points (x_n, y_n, z_n), $n = 1, \ldots, 25$ and their locations on the two images (u_{mn}, v_{mn}), $m = 1, 2$ the Peak Motus system relates them via a Direct Linear Transformation as

$$u_{mn} = \frac{A_m x_n + B_m y_n + C_m z_n + D_m}{E_m x_n + F_m y_n + G_m z_n + 1} \quad (1)$$

$$v_{mn} = \frac{H_m x_n + J_m y_n + K_m z_n + L_m}{E_m x_n + F_m y_n + G_m z_n + 1} \quad (2)$$

Using the 50 equations given by the correspondences, one can determine the 11 parameters for each camera (A_m,..., L_m), via least squares. Knowing the camera parameters, and given a possible coordinate pair of measurements for a (u_1, v_1) and (u_2, v_2),we can write equations (1,2) in terms of unknowns $[x\ y\ z]$:

$$\varepsilon_{\vec{v}} = \begin{bmatrix} A_1 - E_1 u_1 & B_1 - F_1 u_1 & C_1 - G_1 u_1 \\ H_1 - E_1 v_1 & J_1 - F_1 v_1 & K_1 - G_1 v_1 \\ A_2 - E_2 u_2 & B_2 - F_2 u_2 & C_2 - G_2 u_2 \\ H_2 - E_2 v_2 & J_2 - F_2 v_2 & K_2 - G_2 v_2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} D_1 - u_1 \\ L_1 - v_1 \\ D_2 - u_2 \\ L_2 - v_2 \end{bmatrix} \quad (3)$$

Then, this system can be solved using least squares to obtain the 3-D position of a point whose correspondences are known.

To perform accurate reconstruction using the video system, we record video events of bats flying and mark the position of the bat's head in each video frame recorded by the two cameras. Care is taken to ensure that the corresponding points are from the same position on the head.

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 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 each instant the bat emits sonar 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

$$\chi_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2 + (z_i - z_s)^2}. \quad (4)$$

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

$$\chi_i - \chi_j = c t_{ij}. \quad (5)$$

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 [4].

Exact solution For our "L"-shaped microphone array configuration we can employ a novel exact solution [9]. 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

$$\chi_2^2 = r^2 + R^2 - 2rR\cos\theta, \quad \chi_3^2 = r^2 + 4R^2 - 4rR\cos\theta. \quad (6)$$

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 (5). In addition we can write the following identity

$$2\chi_2^2 - \chi_1^2 - \chi_3^2 = -2R^2 \quad (7)$$

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

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

We can get the range and the coordinate x_s as

$$\langle r \rangle = \frac{\chi_1}{2} + \frac{1}{2}\sqrt{\frac{2\chi_2^2 - \chi_3^2 + 2R^2}{2}}, \quad (8)$$

$$x_s = \langle r\cos\Theta \rangle = \frac{3\chi_1^2 - \chi_3^2 + 12R^2}{12R} \quad (9)$$

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. Any errors in these quantities can cause the results to vary wildly or even become imaginary. Further, this exact solution 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 [8]. 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 (5) 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

$$\mathbf{d} = \left[\begin{array}{ccc} \chi_2 - \chi_1, & \cdots, & \chi_M - \chi_1 \end{array} \right]^t. \tag{10}$$

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

$$\chi_i - \chi_j > c_{\min}\left(t_{ij} - \epsilon\right), \quad \chi_i - \chi_j < c_{\max}\left(t_{ij} + \epsilon\right), \tag{11}$$

for $t_{ij} > 0$, and with similar equations for $t_{ij} < 0$. This set of equations (5) and (11) is solved using a constrained L_1 optimization algorithm, termed "CL1" [7]. 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 (10) 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 [6], 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

$$\mathbf{S} = \left[\begin{array}{c} \mathbf{m}_2 - \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_M - \mathbf{m}_1 \end{array} \right], \quad \delta = \left[\begin{array}{c} R_2^2 - d_1^2 \\ \vdots \\ R_M^2 - d_{M-1}^2 \end{array} \right]$$

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

$$\mathbf{S}_W^* \equiv (\mathbf{S}^t\mathbf{W}\mathbf{S})^{-1}\mathbf{S}^t\mathbf{W}, \quad \mathbf{P}_s = \mathbf{S}\mathbf{S}_W^*, \quad \mathbf{P}_S^{\perp} = \mathbf{I} - \mathbf{P}_S \tag{12}$$

$$\chi_1 = \frac{\mathbf{d}^t\left(\mathbf{P}_S^{\perp}\left(\mathbf{W}\left(\mathbf{P}_S^{\perp}\delta\right)\right)\right)}{2\mathbf{d}^t\left(\mathbf{P}_S^{\perp}\left(\mathbf{W}\left(\mathbf{P}_S^{\perp}\mathbf{d}\right)\right)\right)}, \quad \mathbf{x}_s = \frac{1}{2}\mathbf{S}_W^*\left(\delta - 2\chi_1\mathbf{d}\right), \tag{13}$$

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.

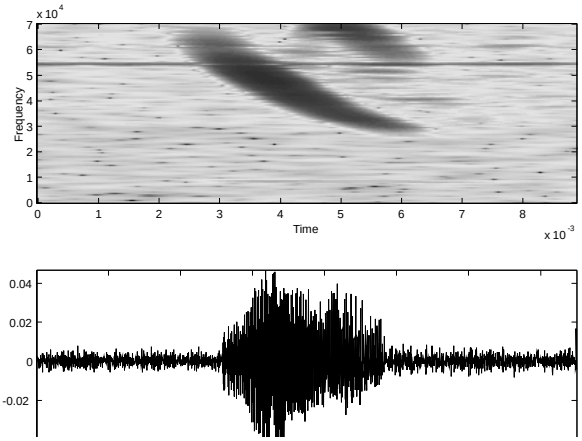

Fig. 2. Spectrogram and waveform of a bat vocalization.

5. RESULTS

We present results from two trials recorded of a bat moving towards a tethered mealworm (insect 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 (~20 Hz). As the bat acquires the target it begins emitting more frequent (~100 HZ) vocalizations (approach). After the capture, indicated by a joining of the estimates of the bat's flight path and the target track, the bat is silent for 200 ms, 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 **d** estimated by CL1 and that from the estimated source position).

In these results there is an offset between the audio and the video estimates of the bat trajectory. In the audio algorithms the microphone locations used were estimated using their video images, and the described DLT algorithm. Since the microphones lie far from the calibrated area, these estimates might be biased, which provides a possible explanation for the discrepancy.

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

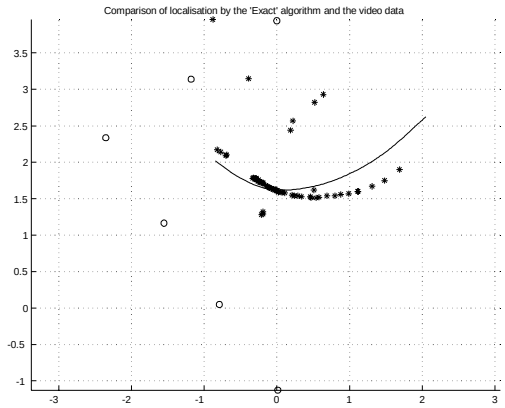

Fig. 3. Exact solution (*) and video data (solid line) for trial 1.

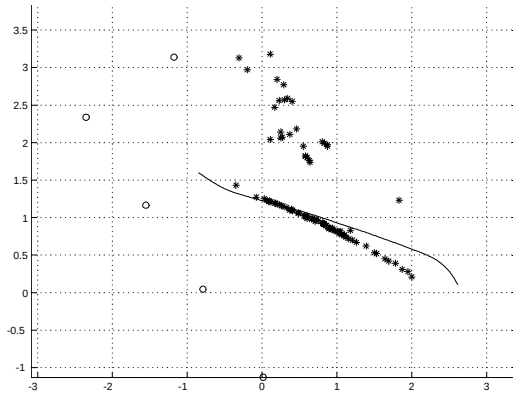

Fig. 5. Exact solution results for the 2nd trial.

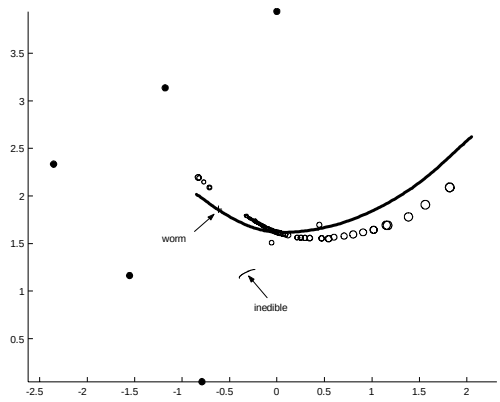

Fig. 4. Comparison of CL1 results and video data. The circles show the positional error estimates of the CL1 results.

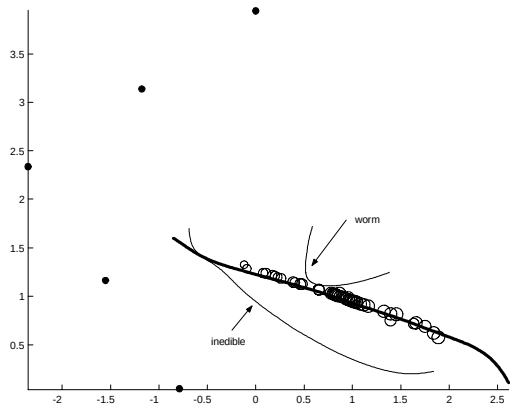

Fig. 6. As in Fig. 4, but now with a moving target and distractor. The bat approaches the correct target, but misses it.

6. CONCLUSIONS

We can make the following preliminary 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 results that compare well with the more complex CL1 algorithm. In addition the exact solution uses fewer closely-spaced microphones, which restricts the cross-correlation lags, resulting in a faster and more robust delay estimate.
- The differences in the video and audio data are larger when the path of the bat and target 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] C.F. Moss & H.-U. Schnitzler (1995) "Behavioral studies of auditory information processing" in A. Popper and R. Fay (Ed) *op. cit.*, pp. 87-145.

[2] D.R. Griffin (1958) "Listening in the Dark", *Yale U.P.*

[3] A. Popper & R. Fay ed. (1995) "Hearing by bats", Springer Handbook of Auditory Research, Vol. 5, Springer.

[4] D.Feitelson & A.Weil (1996). "A robust method for speech signal time-delay estimation in reverberant rooms", *Proc. ICASSP-96, Atlanta, GA.*

[5] R. Hartley & A. Zisserman, (2000) Multiple View Geometry in Computer Vision, *Cambridge*.

[6] 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.

[7] I. Barrodale and F.D.K. Roberts, (1973) "An improved algorithm for discrete l_1 linear approximation," *SIAM J. Numer. Anal.*, **10**, 839-848.

[8] D. Zotkin, R. Duraiswami, L.S. Davis & I. Haritaoglu.(2000) An audio-video front-end for multimedia applications, Proc. IEEE SMC 2000, Nashville, TN.

[9] 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.

[10] C.C. Slama, ed. (1980), Manual of Photogrammetry, 1980. (4th Edition). American Society for Photogrammetry, Falls Church, Virginia.