An Automated Pipeline for 3D Flight Path Reconstruction from Multi-Channel Audio of Echolocating Bats

Sparf Nils Brasar

Master's thesis
2026:E56

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

An Automated Pipeline for 3D Flight Path Reconstruction from Multi-Channel Audio of Echolocating Bats

Sparf Nils Brasar
Supervisor: Karl Åström

June 16, 2026

Abstract

This thesis presents an automated pipeline developed for reconstructing the 3D flight trajectories of bats navigating a cave environment purely from multi-channel audio recordings. The pipeline was developed and validated using the Ushichka dataset which contains multiple simultaneous audio and video recordings of bats in their natural habitat, the Orlova Chuka cave in Bulgaria. The proposed pipeline integrates computer vision, signal processing algorithms, and robust multilateration solvers. Echolocation calls are visually detected from audio spectrograms using a trained SqueezeNet Convolutional Neural Network. Time Differences of Arrival across the microphone array are then estimated via Generalized Cross-Correlation with Phase Transform. To robustly estimate 3D call positions and timing in the presence of anomalous data, a minimal multilateration solver is coupled with a Random Sample Consensus framework. Finally, a multi-hypothesis tracking method associates these discrete coordinates into continuous trajectories. Processing of the acoustic data successfully yielded numerous reconstructed trajectories. Validating the acoustic reconstruction against ground-truth trajectories constructed from thermal cameras, the mean absolute orthogonal distance between them is 3.6 cm. This result demonstrates the usefulness of the approach in studying the flight of bats, and, by extension, their group behavior.

Contents

1 Introduction

The aim of this thesis is to reconstruct the flight paths of bats navigating a cave, purely from audio recordings of their echolocation calls. The method relies on the Time Difference of Arrival (TDOA) of chirps over an ad hoc array of microphones. By measuring the slight time differences of the signal arrival times across different microphones, the original location of the source can be estimated through solving a system of non-linear spatial equations, a process called acoustic multilateration.

The project is based on the Ushichka Dataset, collected by T. Beleyur [1]. The dataset is unique, containing bat calls recorded with an array of ten microphones in their natural habitat over several days. The nature of the Ushichka Dataset requires a novel and robust automated pipeline to be developed in order to solve the well known TDOA problem. A few unique challenges present themselves:

- First, the cave environment - characterized by flat and hard surfaces - produces strong multi-path interference for every bat call, with some echoes being louder than the original signal.
- Secondly, the number of sources present simultaneously - sometimes tens of echolocating bats - introduces ambiguity over which recorded call belongs to which source.
- Furthermore, the sheer volume of the dataset demands that an automated approach be developed with as little manual work as possible to facilitate the ease of use and analysis of the data.

The ultimate goal of T. Beleyur's research is to study the behavior of bats in a group setting, for which accurate estimations of their flight paths and chirp times are an absolute necessity [1].

The report is structured as follows. In Section 2, the theoretical framework for acoustic localization and processing algorithms is established. Section 3 reviews the existing literature regarding the use of CNNs in audio applications and the current state of bat tracking. Section 4 details the design and implementation of the proposed pipeline, with Sections 5 containing the results and discussion. Finally, Section 6 concludes the report and proposes how the method can be further developed.

2 Theoretical Background

2.1 Principles of Acoustic Localization

2.1.1 Trilateration

Acoustic trilateration solves the problem of finding the position $\mathbf{x} \in \mathbb{R}^n$ of a source that emits sound by measuring the Time of Arrival (TOA) of a signal to a set of N receivers. This can be done when the locations of the receivers are known and the source and receivers are synchronized so that the absolute distance between the source and each receiver can be inferred by the TOA and the speed of sound. More formally, when a signal is emitted at time t_0, the TOA to the receiver $j = 1, 2, 3..., N$ is measured as t_j. If we then take the difference of emission and TOA and multiply it by the speed of sound v, we obtain the inferred distance $d_j = v(t_j - t_0)$. Then let us form the Euclidean norm of the difference $||\mathbf{x} - \mathbf{r_j}||$ between the unknown position $\mathbf{x}$ of the source and the known position $\mathbf{r_j}$ of receiver j. Estimating the location of the source in a Maximum Likelihood (ML) way when assuming independent Gaussian noise comes down to solving the following minimization problem [2]

$$\arg\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{j=1}^{N} (||\mathbf{x} - \mathbf{r_j}|| - d_j)^2. \tag{1}$$

The minimal case for the 3D trilateration problem is one source and three receivers. Each TOA constricts the possible source locations to a sphere. If we assume that $||\mathbf{x} - \mathbf{r_j}|| \approx d_j$ for every j then for two receivers, the range spheres of possible source locations intersect in a circle. Adding a third receiver the three spheres intersect in at most two possible points. These two solutions are both mathematically valid, but often other logic can be employed to eliminate one of them. If not, a fourth receiver should restrict the possible solution to one of them. See Figure 1 for an example illustrating the minimal case from 2D space. The range spheres are circles in 2D space.

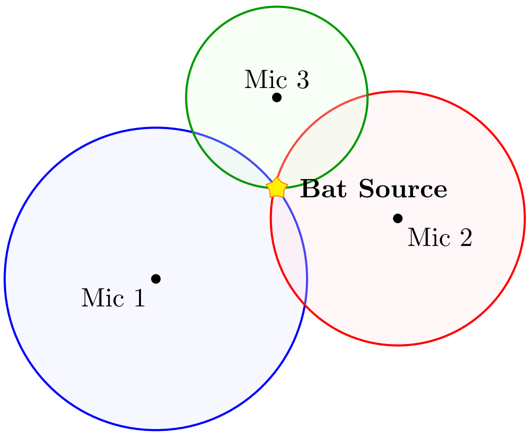


Figure 1: 2D Trilateration: The intersection of two range circles occurs at two points. A third range circle is required to identify the correct source position.

2.1.2 Multilateration

Trilateration hinges on the fact that the source and receivers share a synchronized clock to obtain the inferred distances d_j. In the case of passive sensing of an uncontrolled source this is not immediately possible. This method is instead called Multilateration. Forming the difference between the emission time t_0 and the TOA t_j of receiver j and multiplying by the speed of sound v, we obtain $v(t_j - t_0) = v \cdot t_j - v \cdot t_0 = d_j - o$. Note that d_j here refers to a fundamentally different quantity than the inferred distance in trilateration. Because the time of emission t_0 is unknown we have introduced a fourth unknown variable o we call the offset. Again form the Euclidean norm of the difference $||\mathbf{x} - \mathbf{r_j}||$ and obtain the equation

$$||\mathbf{x} - \mathbf{r_j}|| = d_j - o. \tag{2}$$

For receivers $j = 1, 2, 3..., N$ we can construct N equations, giving us

$$\begin{cases} \|\mathbf{x} - \mathbf{r}_1\| = d_1 - o \\ \|\mathbf{x} - \mathbf{r}_2\| = d_2 - o \\ \vdots \\ \|\mathbf{x} - \mathbf{r}_N\| = d_N - o. \end{cases} \tag{3}$$

The system is difficult to solve due to the norm, so to eliminate it we can start by squaring both sides,

$$\begin{cases} \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{r}_1 - \mathbf{r}_1^T\mathbf{x} + \mathbf{r}_1^T\mathbf{r}_1 = d_1^2 - 2d_1o + o^2 \\ \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{r}_2 - \mathbf{r}_2^T\mathbf{x} + \mathbf{r}_2^T\mathbf{r}_2 = d_2^2 - 2d_2o + o^2 \\ \vdots \\ \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{r}_N - \mathbf{r}_N^T\mathbf{x} + \mathbf{r}_N^T\mathbf{r}_N = d_N^2 - 2d_No + o^2 \end{cases} \tag{4}$$

followed by subtracting every other equation with the first one to eliminate the squared terms of $\mathbf{x}$ and o, resulting in the following, [3].

$$\begin{cases} \mathbf{x}^T\mathbf{r}_1 - \mathbf{x}^T\mathbf{r}_2 + \mathbf{r}_1^T\mathbf{x} - \mathbf{r}_2^T\mathbf{x} + \mathbf{r}_2^T\mathbf{r}_2 - \mathbf{r}_1^T\mathbf{r}_1 = d_2^2 - d_1^2 - 2d_2o + 2d_1o \\ \mathbf{x}^T\mathbf{r}_1 - \mathbf{x}^T\mathbf{r}_3 + \mathbf{r}_1^T\mathbf{x} - \mathbf{r}_3^T\mathbf{x} + \mathbf{r}_3^T\mathbf{r}_3 - \mathbf{r}_1^T\mathbf{r}_1 = d_3^2 - d_1^2 - 2d_3o + 2d_1o \\ \vdots \\ \mathbf{x}^T\mathbf{r}_1 - \mathbf{x}^T\mathbf{r}_N + \mathbf{r}_1^T\mathbf{x} - \mathbf{r}_N^T\mathbf{x} + \mathbf{r}_N^T\mathbf{r}_N - \mathbf{r}_1^T\mathbf{r}_1 = d_N^2 - d_1^2 - 2d_No + 2d_1o. \end{cases} \tag{5}$$

Gathering the terms of $\mathbf{x}$ and o gives us

$$\begin{cases} 2(\mathbf{r}_1 - \mathbf{r}_2)^T\mathbf{x} + 2(d_2 - d_1)o = d_2^2 - d_1^2 - \mathbf{r}_2^T\mathbf{r}_2 + \mathbf{r}_1^T\mathbf{r}_1 \\ 2(\mathbf{r}_1 - \mathbf{r}_3)^T\mathbf{x} + 2(d_3 - d_1)o = d_3^2 - d_1^2 - \mathbf{r}_3^T\mathbf{r}_3 + \mathbf{r}_1^T\mathbf{r}_1 \\ \vdots \\ 2(\mathbf{r}_1 - \mathbf{r}_N)^T\mathbf{x} + 2(d_N - d_1)o = d_N^2 - d_1^2 - \mathbf{r}_N^T\mathbf{r}_N + \mathbf{r}_1^T\mathbf{r}_1, \end{cases} \tag{6}$$

which is a system of equations that is linear in our unknowns $\mathbf{x}$ and o. To further simplify, we can choose receiver 1 as a reference and set the TOA $t_1 = 0$. This effectively means that we can reinterpret the other TOAs t_j as the difference in arrival time compared to the reference receiver. Moreover, the magnitude of the offset o can then also be reinterpreted as the distance from the source to the reference receiver. Thus we can solve the Multilateration problem, not by measuring the TOAs, but instead by measuring the Time Difference of Arrival (TDOA) to every receiver from a selected reference, giving us the final system of equations to locate the source.

$$\begin{cases} 2(\mathbf{r}_1 - \mathbf{r}_2)^T\mathbf{x} + 2d_2o = d_2^2 - \mathbf{r}_2^T\mathbf{r}_2 + \mathbf{r}_1^T\mathbf{r}_1 \\ 2(\mathbf{r}_1 - \mathbf{r}_3)^T\mathbf{x} + 2d_3o = d_3^2 - \mathbf{r}_3^T\mathbf{r}_3 + \mathbf{r}_1^T\mathbf{r}_1 \\ \vdots \\ 2(\mathbf{r}_1 - \mathbf{r}_N)^T\mathbf{x} + 2d_No = d_N^2 - \mathbf{r}_N^T\mathbf{r}_N + \mathbf{r}_1^T\mathbf{r}_1, \end{cases} \tag{7}$$

Because the number of receivers N is often large, this system will be overdetermined. A least squares approach is possible, but likely not desirable as is discussed further in Section 2.3 about RANSAC.

Due to the addition of a fourth unknown, the offset o, the minimal case to solve the multilateration problem in 3D is one source together with four receivers. Note that by subtracting the squared reference equation $\|\mathbf{x} - \mathbf{r}_1\|^2 = o^2$ from the others, we reduced the above problem to a system of $N-1$ linear equations. This means that when solving the minimal problem, we use $4 - 1 = 3$ linear equations, which is just enough to express $\mathbf{x}$ in terms of o. To definitively solve the system, we must substitute $\mathbf{x}$ expressed in terms of o into the reference equation $\|\mathbf{x} - \mathbf{r}_1\|^2 = o^2$. Because of the norm, we will typically obtain two mathematically valid solutions for the pair $(\mathbf{x}, o)$.

Another difference compared to the trilateration problem is the geometric interpretation. In trilateration each TOA restricts the possible source locations to a range sphere, whereas in multilateration it is the TDOAs between pairs of microphones that restrict the solutions. The TDOA between two microphones is proportional to the difference in distance, meaning that every point in 3D space that preserves the difference in distance to two focal points (the receivers) is a possible source location. This no longer creates a range sphere, but one branch of a hyperboloid depending on the sign of the TDOA. In 2D, each TDOA produces a range hyperbola, as shown in figure 2. While the intersection of the hyperboloids in 3D space can result in a unique solution, two solutions are common. Similarly to trilateration, a priori knowledge or a fifth receiver can disambiguate if two mathematically valid solutions are produced [4].

2.2 GCC-PHAT

A widely used method for estimating the TDOA of a signal is the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [5]. See Figure 3 for an example. This framework, first developed by Knapp and Carter [6], involves prefiltering a pair of received signals, before they are passed through a cross-correlator. If we let $X_1(f)$ and $X_2(f)$ denote the Fourier transform of two unfiltered signals, the cross power spectral density function $G(f)$ is defined as

$$G(f) = X_1(f)X_2^*(f), \tag{8}$$

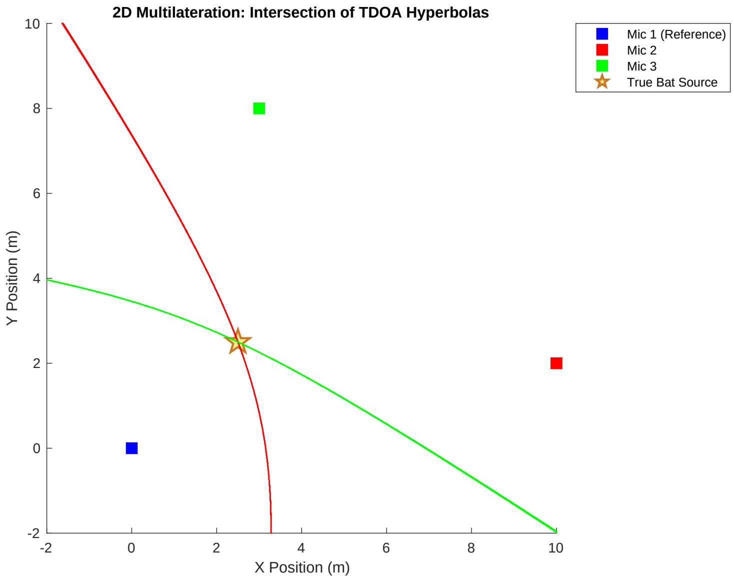


Figure 2: 2D multilateration. The TDOA between the reference microphone (blue) and each of the other microphones restricts the possible source locations to a hyperbola. The source is found in the intersection of the hyperbolas.

where * denotes the complex conjugate. For GCC-PHAT the chosen prefilter is the Phase Transform (PHAT), which weights the cross power spectrum by the inverse of its own magnitude at each frequency. This effectively normalizes the influence across all frequencies, leaving only the phase shift. After prefiltering, the cross-correlation of the two signals is calculated for different time-delays. The cross-correlation is related to the cross power spectrum by the inverse Fourier transform, which leads us to the final expression for the GCC-PHAT function,

$$R_{PHAT}(\tau) = \int_{-\infty}^{\infty} \frac{X_1(f)X_2^*(f)}{|X_1(f)X_2^*(f)|} e^{i2\pi f\tau} df, \tag{9}$$

where τ represents the time delay. The specific delay $\hat{\tau}$ that maximizes the cross-correlation is identified as the real time delay between the two signals. Thus, the estimation of the TDOA between the two signals is

$$\hat{\tau} = \arg\max_{\tau} R_{PHAT}(\tau). \tag{10}$$

While the PHAT filter is described as "ad hoc", it achieves the highly desirable effect of producing very sharp correlator peaks compared to other prefilters. This allows for exact TDOA estimations to be made. Furthermore, PHAT handles reverberations comparatively well, an advantage when working with signals produced in acoustically reflective environments [5].

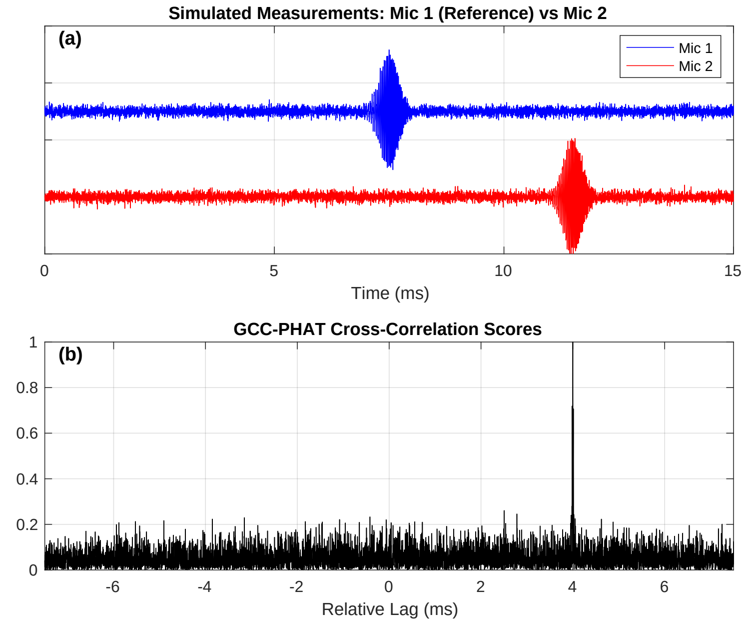


Figure 3: TDOA estimation via GCC-PHAT. (a) Two recordings of the same simulated signal. The blue (top) signal from Mic 1 is used as the reference for the algorithm. (b) Normalized cross-correlation magnitude of the blue (top) and red (bottom) signal, showing a sharp peak at 4 ms, correctly identifying the TDOA between the red signal and the blue reference.

2.3 Robust Estimation with RANSAC

Random Sample Consensus (RANSAC), developed by Fischler and Bolles [7], is a robust estimation algorithm designed to fit a mathematical model to a dataset in a way such that noise and error in the data minimally affect the outcome of the final model. Compared to other smoothing and fitting methods such as the Least Squares method, which attempts to minimize the sum of squared residuals over all data points, and therefore can be heavily biased by outliers in the data, RANSAC tests multiple hypotheses to find the model that best explains the most data.

The process begins with randomly selecting a minimal subset of data points to estimate all the model parameters. The remaining data points are then compared to the resulting model, and the errors are calculated. If an error is within a certain tolerance that point is counted as an inlier, and if not it is labeled as an outlier and ignored. After iterating over a fixed number of random selections, the model that best explains the data, i.e. the one with the highest number of inliers, is selected. Finally, an optimization step involving the inliers is usually performed . An example of the usefulness of RANSAC is for the problem of calculating the position of a bat from the TDOAs over a microphone array. In some of the microphones the TDOAs might be poorly calculated due to multi-path interference, or the signal was not measured adequately. Using a traditional smoothing technique like Least Squares, the result would be skewed by the outliers, but when RANSAC is employed, the most successful hypothesis survives while managing to ignore bad data.

A simple example to illustrate how RANSAC handles errors in a dataset is fitting a straight line to a set of points. In Figure 4 the dashed red line follows the Least Squares fit compared to the blue one produced by RANSAC.

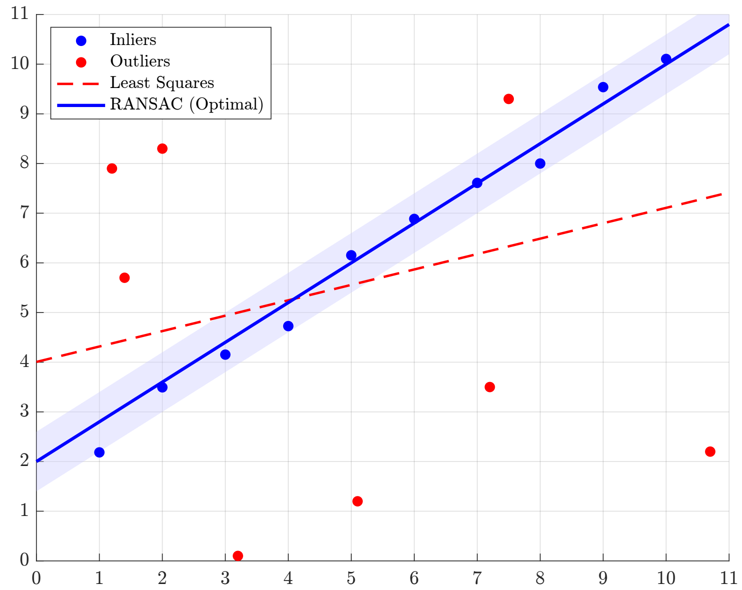


Figure 4: Example of difference between Least Squares (LS) and RANSAC. LS is heavily biased by outliers, while RANSAC manages to identify the underlying relationship. The semi-transparent blue region represent the tolerance-distance for a point to be counted as an inlier.

3 Literature Review

To motivate the architecture of the proposed pipeline, this section reviews the use of CNNs in acoustic applications together with the history and current state of 3D bat tracking. By examining existing methods and their limitations, this review identifies the design inspirations that influenced the final pipeline.

3.1 Convolutional Neural Networks (CNNs) for Audio

Although Convolutional Neural Networks (CNNs) were developed for computer vision, they have proven highly successful in classifying audio data when the signal is represented by a spectrogram, [8]. By converting audio to a 2D time-frequency representation via the Short-Time Fourier Transform (STFT), CNNs can treat acoustic features such as the short frequency sweeps of bat calls as visually identifiable patterns.

Hershey et al. [9] evaluated and compared several CNN architectures, among them AlexNet, on a large audio dataset, demonstrating that they significantly outperform fully connected networks. Similarly, Eutizi et al. [10] describe using the CNN SqueezeNet for classifying audio events. SqueezeNet achieved a validation accuracy of 89.93% across ten classes of urban sounds.

SqueezeNet was specifically developed in 2016 by Iandola et al. [11] to be a highly accurate CNN with a very small size. The group presented results of AlexNet-level accuracy with 50 times fewer parameters. The small size of SqueezeNet brings with it distinct advantages for this project in training and deployment.

3.2 Acoustic Localization of Bats

Since the formal discovery of the ultra-sonic echolocating calls of bats by Pierce and Griffin [12] in 1938, the study of bats has evolved significantly. The first example of the specific use of a microphone array to infer the positions of bats was by Roeder [13] already in 1966. Subsequent advances in various fields have allowed bats to be studied by their calls in more sophisticated ways, specifically using large microphone arrays and TDOA multilateration [14].

An example of using a microphone array to reconstruct the flight path of bats was given by de Framond et al. [15] in 2023. The group caught and released bats individually and let them fly over a rigid planar array (∼1 m in diameter) consisting of four microphones. In total the paths of five bats were reconstructed using a custom-made software written by Peter Stilz. To reconstruct the flight paths the group first filtered the audio with a band-pass filter (20-90 kHz) and used a fixed threshold to detect calls by amplitude from a central microphone in the array. Then, the TDOA to the peripheral microphones were calculated through cross-correlation, although the specific design used is not elaborated on. With the TDOAs available an estimation of the position was calculated and manually controlled and combined with other position estimates to form a trajectory. Outlier positions were manually excluded and calls that were outside of a specific duration window (1-8 ms) were automatically excluded. This resulted in a total of 98 calls used for path reconstruction. The accuracy of the method was evaluated using simulated calls transmitted by a speaker $3-10$ m away from the array. The evaluated quantity was the percentage difference in distance to the central microphone between the ground truth position and the estimate. While the overall difference did not show a significant bias with a mean error of -0.5% alongside a standard deviation of $\sigma = 4.1\%$, the error range across the Y-axis was the largest ranging from $-25.5\% - 11.2\%$. Because the authors evaluated their errors using a Gaussian Generalized Linear Model, we can estimate the Mean Absolute Error (MAE) of the difference in real versus estimated distance using the standard deviation. For a normally distributed variable the MAE relates to the standard deviation σ according to the well known approximate relationship $MAE \approx 0.8\sigma$. A σ of 4.1% would correspond to an MAE of around 3.3%, which over the tested range of $3-10$ implies a range of MAEs in distance from the central microphone of $0.1-0.3$ m.

Another example is the use of a rigid small scale array of 64 microphones (1.2×1.2 m) by Verreycken et al. [16] in 2021 to study four individual bats hunting in a flight-room. They used a TDOA approach outlined in [17] to perform probabilistic localization estimates. Call detections were made by a matched filter to identify parts with the desired features and were then used to produce TDOA estimates.

While de Framond et al. [15] used filtering and a threshold to detect bat calls, Mac Aodha et al. [18] in 2018 opted for a CNN to visually detect calls from spectrograms. The group developed an open source pipeline called Bat Detective designed to monitor bat populations in outdoor environments. The pipeline begins with filtering and denoising audio recordings to produce a clean spectrogram. A custom-built CNN trained on similar recordings was then applied to the spectrograms in a sliding window fashion, to determine the probability that a chirp is present at each point in time. Finally, the probability peaks were condensed into a single point in time, locating the calls precisely. The group found that their CNN outperformed all other applications designed to detect bat calls on available test sets, with a maximum average precision of 86.3% and a maximum recall rate of 81.8% at 0.95 precision. The training set used for the CNN was produced by letting volunteers manually annotate spectrograms. In the end, the labels produced by a single volunteer were used for both training and testing, with $4,782$ instances of call annotations with an accuracy of 84.5%.

These examples demonstrate the current state of bat tracking through audio. The number of bats being studied is often small (4-5 individuals), either capturing and releasing them in their natural habitat or in a controlled environment. The arrays used are rigid and often small-scale, limiting the area where acoustic multilateration can be successfully performed [17]. In terms of analyzing the recordings and performing reconstruction estimates, current methods rely on a high degree of manual labor. The recording environments are often acoustically controlled [17] or outdoors [15] [18], limiting the potential issue of multi-path interference.

The Ushichka dataset outlined in the next section represents a qualitative and quantitative step forward in the field [1], and requires a more robust and streamlined analyzing approach be developed to handle the unique challenges it posits.

4 Methodology

4.1 Implementation specifications

The entire software pipeline was implemented in MATLAB R2023b. All computations and simulations were performed on a MacBook Air (2021) equipped with an Apple M1 chip and 8GB of RAM.

4.2 The Ushichka Dataset

The Ushichka dataset was collected by T. Beleyur [1] in the summer of 2018 in the Orlova Chuka cave system in Bulgaria. It contains simultaneous audio recordings from a large-scale 10 microphone array, thermal video recordings from 3 cameras, LiDAR scans of the cave, and weather data. For a picture of the cave and the sensor set up see Figure 5. The bats present in the cave are of various different species, but the two species that were the focus of the Ushichka dataset are *Myotis myotis* and *Myotis blythii*. They are morphologically similar and produce indistinguishable calls.

Figure 5: The experimental set up in the Orlova Chuka cave. The black inverted T in the center of the image holds four microphones. Six smaller microphones are mounted on the left wall and not visible in this photograph.

The basis for this project is the multi-channel audio recorded by the array, specifically the audio from the night of June 21-22. The audio was sampled at 192 kHz. Alongside the audio, the thermal video recordings were used to evaluate the accuracy of the reconstructed paths. The weather data, consisting of humidity and temperature measurements are important for accurately estimating the speed of sound, a crucial parameter for accurate TDOA estimation. The microphone set up consisted of two large-scale arrays (maximum distance between microphones is $\sim$5.5 m), one of them being a rigid inverted T configuration of four microphones and the other a W configuration consisting of six microphones freely mounted on a cave wall. A completely foundational requirement for TDOA multilateration is that the locations of the receivers are known. For the Ushichka dataset this necessitates a robust calibration of all microphones, but especially the freely mounted ones. This calibration is laid out in Section 4.2 and while it is an extra step, the method allows for freely mounted arrays to be more readily used, facilitating field work and location-specific set ups.

4.3 Microphone Array Calibration

A method for self-localization of microphones in an array using ambient sound is presented by Zhayida et al. in [19]. The method relies on estimating TDOAs of sound from a moving source using GCC-PHAT, with the localizations being robustly estimated with RANSAC. In order to estimate the parameters with RANSAC one can use one of three minimal solvers detailed in [20] by Kuang and Åström.

This approach was refined for the specific case of locating the microphones used to produce the Ushichka dataset. The method is described in [21] by Batstone et al. In order to calibrate the microphones, a synthetic sound was emitted with regular known intervals and moved around in the recording volume. While the emission times are not known, the known interval between each emission places the problem somewhere between a TOA and pure TDOA problem. To obtain the TDOAs a matched filter was used to detect where in each audio file the synthetic signal was present. The minimal solver used involved 5 senders and 5 receivers to produce 4 possible solutions. To robustly estimate the microphone positions RANSAC was used with the minimal solver. When the same method was used in an office experiment the differences in Euclidean distance between the estimated and ground-truth positions were in the range of 0.0587-0.2110 m.

4.4 Proposed Pipeline

The proposed workflow from a multi-channel audio recording to reconstructed flight paths can be divided into three broad steps.

1. Automatic call detection - The spectrograms of all audio channels are produced and scanned by a trained SqueezeNet model, sliding a window over the entire duration of the recording. This results in certain time stamps being labeled as a call and passed on to the subsequent steps.

2. TDOA multilateration with RANSAC - Every call detection is evaluated by cutting out a section around the corresponding time stamp in the raw audio file, obtaining the TDOAs with GCC-PHAT. The minimal problem is solved with RANSAC, picking a minimal subset of microphones to estimate the position and controlling the number of inliers. The solution with the highest number of inlier microphones is picked as the position estimate and passed on to the next step.

3. Trajectory association and hypothesis tracking - An ad hoc method to piece together individual position estimates into flight trajectories was developed. It keeps track of multiple hypotheses, evaluating whether a position can be connected to an existing path through a number of

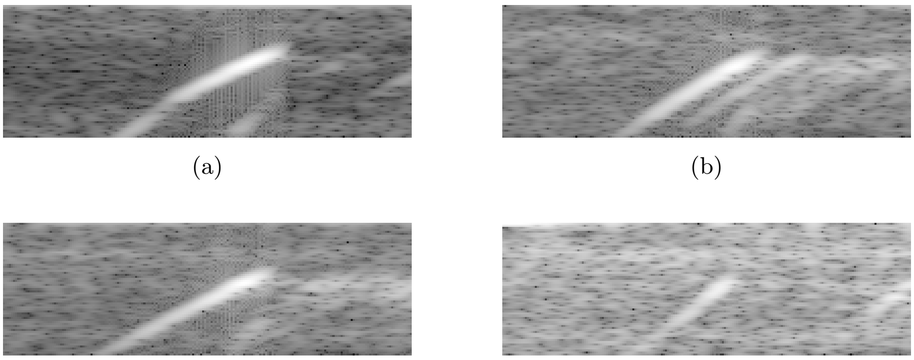
physical criteria, including distance in time and space, speed, acceleration, and turn angle.

Each step is elaborated on in their respective subsections. From here on, we will call the SqueezeNet model trained for this project BatNet. The training of BatNet was performed prior to deploying the pipeline and is described in the following subsection.

All spectrograms were generated in the same manner using the Short Time Fourier Transform (STFT) with a window size of 128, an overlap of 120 and a FFT window of 128, resulting in 65 frequency bins. In order to improve visibility of fainter signals the absolute value followed by a logarithm was applied to the spectrograms.

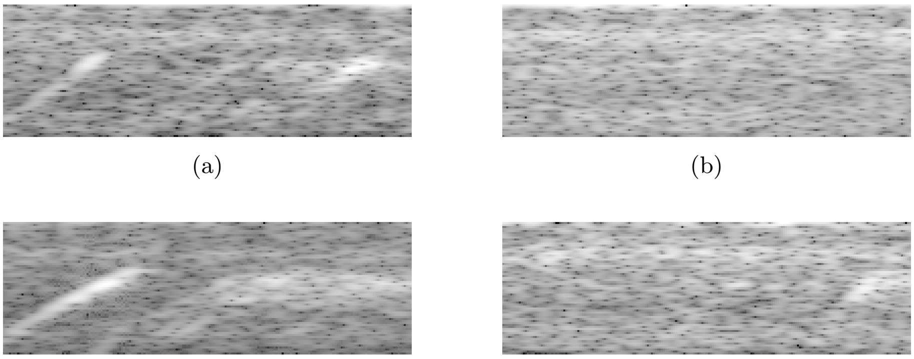
4.4.1 Training the Network

The specific way the training set was created ultimately depended on the challenges posed by multi-path interference (echoes) and the strategy chosen to detect calls, which was a sliding window approach. The training set consists of images cut out from spectrograms from the Ushichka dataset. Each image encompassed all available frequencies and had a size of 65 frequency bins by 201 time frames, but were stretched to fit the required input size of 227×227 of SqueezeNet. The time window of 201 spectrogram pixels corresponds to around 9 ms of audio which is enough to capture a call and eventual multi-path interference. The images in the set were manually labeled by the author as one of the categories "Call" or "Non-Call", with 27 and 89 total images respectively. See figures 6 and 7 for some examples of images from the two categories. Note that the frequency axis is flipped in the images; this, however should have no bearing on how SqueezeNet interprets the images. An image is labeled a call if it has a call-like feature in the middle of the image, possibly with call-like features to the right, but never with call-like features to the left, as this would indicate that the image is centered on an echo.



(a) (b) (c) (d)

Figure 6: Examples of images from the training set labeled as "Call". Note the prominent echoes to the right of the main call in (b). The calls appear more clearly in (a) and (c) compared to (d).



(a) (b) (c) (d)

Figure 7: Examples of images from the training set labeled "Non-Call". (a) is centered on random noise in between two louder features. (b) is lacks any call-like features. (c) is centered on echoes after a real call. (d) is centered on random noise, placing the call-like feature to the right in the image.

Due to the small size of the training set, the data was augmented during training to prevent overfitting. This was done by stretching and squashing the images as well as slightly translating up and down, and right and left. Other specific settings were the minibatch size of 16, learning rate of 0.001, and a number of 15 epochs. To evaluate the performance of BatNet on a test set, precision (the percentage of predicted calls that were actually calls) and recall (the percentage of true calls that were successfully predicted as calls) were calculated.

4.4.2 Automatic Call Detection

To detect calls in the audio, a .wav file is first converted to spectrograms. BatNet was then used to scan the spectrogram in a sliding window fashion. Each window encompassed 65×201 pixels, representing all available frequencies and roughly 9 ms of audio. The input size of the SqueezeNet architecture is strictly 227×227 pixels, so the spectrogram images were resized and interpolated for compatibility. In order to save time, the scan was performed in two rounds, first using a probing coarse scan followed by a thorough fine search. The coarse scan identified regions of interest by taking strides of 150 pixels between windows. Any regions with a probability of containing a call above 10% was also investigated with the fine scan, taking strides of 10 pixels. The fine scan created probability peaks, and any region with a probability of 95% or higher of being a call was deemed to actually contain one. In order to extract an exact time index for a call, all frequency amplitudes for a specific time frame were added together. The maximum total amplitude over a search window that was slightly extended around the probability peak was selected as the call index. See Figure 8 for an example of a detection.

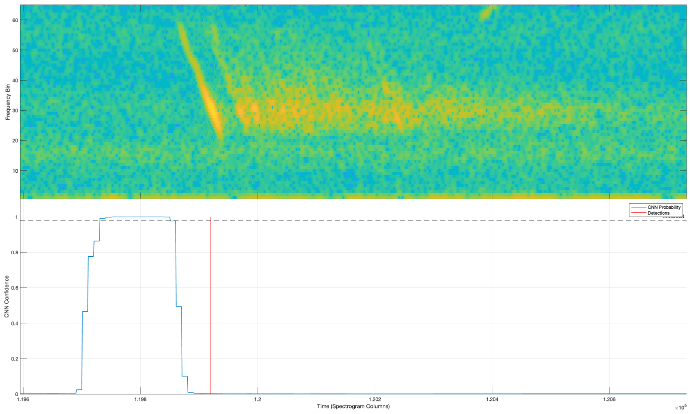


Figure 8: An example of an automated call detection. The top window contains the spectrogram of a bat call. The bottom window contains the probability curve from BatNet in blue, the call detection threshold is a dashed grey line, with the final call detection being marked by the vertical red line.

4.4.3 TDOA Multilateration with RANSAC

After obtaining the specific time indices for call detections the localization of the emitting bat can begin. Every single detection was handled in the following way. First, a reference slice around the detection was cut out from the relevant audio channel. The reference window encapsulates 10 ms of audio. GCC-PHAT was then used to extract the TDOA of the signal compared to every other channel. The slice cut out from the other channels is centered around the time of the reference call, but wider to account for the time it takes for the sound to reach every microphone. The size of the slice is based on the maximum distance between to microphones and encapsulates 26.4 ms of audio. The reference audio slice was symmetrically padded with zeros to make it the same size as the other slices, which the GCC-PHAT implementation requires.

The TDOAs are then handed to a minimal solver coupled with a RANSAC framework. A minimal subset of microphones and TDOAs were selected to calculate the bat position and time offset. RANSAC is usually designed to pick a set number of random minimal subsets, but in this implementation the number of unique subsets of four microphones is only $\binom{10}{4} = 210$, so an exhaustive search of all possible subsets was made. Using the estimated position and time offset, the resulting TDOAs over the microphone array compared to a reference microphone was calculated and compared to the measured TDOAs. Any microphone with a difference between measured and estimated TDOA within $\frac{0.02 \text{ m}}{v} \approx 58.8\ \mu s$, where v represents the local speed of sound, was considered an inlier. The residuals were calculated as the difference between the estimated and measured TDOAs for the inliers. The solution with highest number of inliers was selected as the best estimate. If multiple solutions had the same number of inliers, the one of them with the lowest Euclidean norm of the residuals was selected. After this, the solution is optimized to best fit all the inliers using a modified Gauss-Newton approach.

4.4.4 Trajectory Association and Hypothesis Tracking

When all call detections have been processed, individual estimated positions need to be connected to form trajectories. Because the same emitted call can be processed up to ten times if detected in every channel, position estimates often cluster together. Position estimates within certain time (< 10 ms) and distance (< 15 cm) limits were considered to be instances of the same emission position and the point with the smallest Euclidean norm of the RANSAC residuals was selected as the best estimate of the group. The other estimates were discarded. The position estimates were then looped through in chronological order to build the trajectories. The creation of the trajectories is done by considering multiple hypotheses simultaneously and selecting the most successful hypothesis in the final step.

We consider appending a position estimate when a few trajectory hypotheses have already been constructed. The position estimate is deemed appendable to a trajectory hypothesis if a set of physical criteria are met. If the point can be appended to a hypothesis, this creates two competing hypotheses: the old one without the new point and the new one with the position estimate added. During this stage, every position estimate can be part of multiple hypotheses as long as the physical criteria are met. A hypothesis is also considered where the new position estimate is the start of a new trajectory.

The first physical criterion is a time limit; if the time difference between the last position estimate and the next considered one is larger than 25 ms it can not be appended, and the path hypothesis is deactivated since no subsequent position estimates will be appendable either. A maximum distance of 1 m between two points is also enforced. If a bat is required to fly with a speed greater than 7 m/s to reach the next position, then it is not appended. For every path hypothesis consisting of two or more points, the maximum magnitude of the acceleration and turn angle required to reach the new point are calculated. The physical limits chosen here are 30 m/s^2 for the acceleration magnitude and 50° for the turn angle. After all position estimates have been considered, the competing hypotheses are resolved. This is done by considering the hypotheses by the number of points of which they consist in decreasing order. Every point in the path is labeled as claimed, and if it is part of a shorter path, the claimed status of one of its constituent points discards the entire shorter hypothesis. Finally, all paths consisting of fewer than 6 points are discarded. The trajectories were then visualized using a piecewise cubic Hermite interpolation between each individual call position.

4.5 Synthesis of Multi-Target Audio

To further evaluate the ability of the pipeline to track multiple moving sources, audio of individual bats was synthesized into multi-bat audio. An audio recording containing three separate instances of single bats was used to produce the synthesized audio. The three instances were overlaid by extracting the relevant parts and adding the raw audio waveforms together.

This synthesized audio effectively represents three bats simultaneously echolocating. The synthesized audio was then passed through the entire pipeline.

5 Results and Discussion

5.1 Performance of SqueezeNet

The trained SqueezeNet model BatNet was evaluated on a separate set of images consisting of 277 entries, obtaining a recall of 96.6% and a precision of 72.3% using a classification threshold of 0.5. See Figure 9 for a Confusion matrix of the test set. Comparing this result to the one achieved by Mac Aodha et al. [18] of a recall of 81.8% at a precision of 95%, we see that BatNet obtained a significantly higher recall, albeit with a lower precision. The trade off between precision and recall is difficult to navigate, but for the purposes of this project, false negatives (missed true calls) are more costly than false positives (false alarms); a false positive only requires more computing power and will reliably be discarded later by the path association algorithm, whereas a false negative misses a potentially crucial call needed to reconstruct a trajectory.

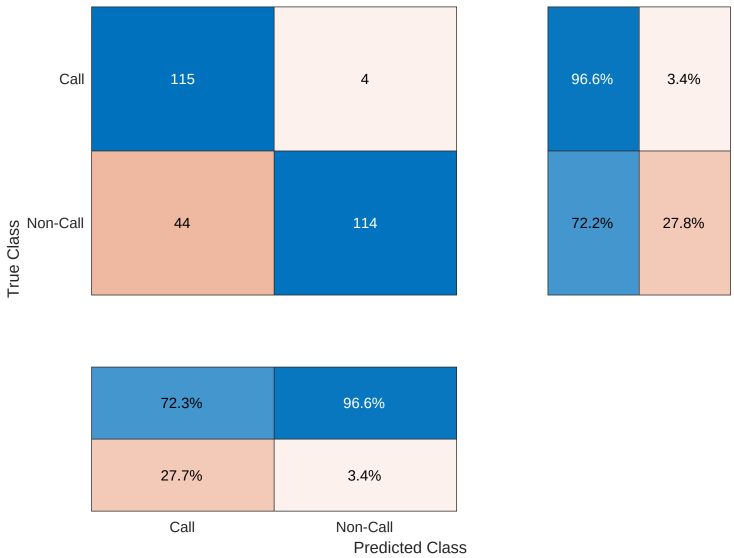


Figure 9: The Confusion matrix of BatNet applied to the test set. The top left figure displays the absolute number of images pertaining to their true classes and what prediction they were given by SqueezeNet.

The small size of the training set is a notable detail. Producing high quality training data is a challenge, and it is not immediately clear what constitutes as an optimal training example. Attempts have been made to create a more accurate model using a larger training set, but the BatNet model was never outperformed. The question of why this is remains unanswered, but it is possible that a larger training set introduced less accurate images. However, the fact that accurate results could be obtained with such a rudimentary approach demonstrates the usefulness of the method. Ultimately, for the performance of BatNet to be rigorously evaluated, the resulting detections would need to be investigated by an expert in bat vocalizations, which remains a subject for future work.

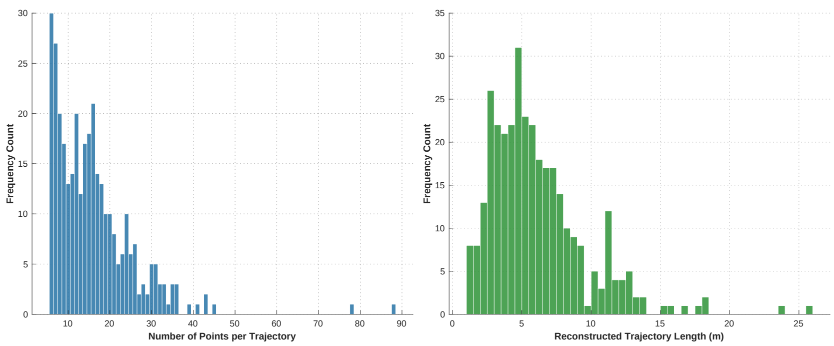
5.2 Reconstructed trajectories

A total of 13 separate recordings encapsulating 533.1 s of audio were analyzed using the pipeline. Across all 13 files and 10 recording channels, a total of 62,865 calls were detected, with 5,376 individual detections used to reconstruct around 330 trajectories. The continuity of the reconstructed trajectories varied widely with some only lasting a few tenths of a second, while others captured a bat circling the space for long periods and executing large maneuvers. The physical parameters of the reconstructed trajectories are aggregated in Table 1. The number of points in the trajectories ranged from $6 - 88$ with an average of 16. The lengths ranged from $1.02 - 25.52$ m with an average length of 6.12 m. The temporal duration of the trajectories ranged from $0.35 - 7.20$ s lasting an average of 1.56 s.

Table 1: Summary of acoustically reconstructed trajectories ($N = 335$)

Metric Parameter	**Minimum**	**Average**	**Maximum**
Physical trajectory length (m)	1.02	6.12	25.52
Temporal flight duration (s)	0.35	1.56	7.20
Number of points	6	16.0	88

The lengths and number of points of all the trajectories are gathered in two histograms in Figure 10. The length of paths are mostly clustered between 1-10 m, with a smaller number of paths exceeding 10 m. However, the lengths of the trajectories are limited by the size of the cave chamber which had dimensions of $5 \times 3 \times 1.6$ m^3. Consequently, an average reconstructed trajectory length of 6.12 m likely captures a large part of the true flight paths in the recording volume.



(a) (b)

Figure 10: In (a) the number of points per trajectory are visualized with a histogram. Each bin represents a unique integer. In (b) the lengths of all trajectories are visualized with a histogram. Each bin captures a width of 0.5 m.

A major success of the pipeline is its ability to track single bats for extended periods of time over the entire recording space. The longest reconstructed trajectory of 25.52 m, shown in Figure 11a follows a bat as it enters the recording volume, loops around the central chamber, and finally exits through another opening. The bat was successfully tracked throughout the entire duration that it was present in the recording volume, showing that every part of the pipeline performed as envisioned. It is of course an outlier, but even so it represents a qualitative and quantitative step forward in the field. Another feature of the pipeline that shows its robustness is its ability to track multiple bats simultaneously, as shown in Figure 11b. The pipeline was able to detect, estimate the position, and associate simultaneous calls belonging to up to five different individuals. Across all reconstructed trajectories (N = 335), there were several instances of up to five bats being tracked simultaneously. Table 2 summarizes the number of instances and the cumulative time spent tracking groups by number. To demonstrate the capacity to handle volumes of calls, figures 11c and 11d display the 51 reconstructed trajectories extracted from a single 50 s audio recording. In addition to showing the potential for handling large quantities of data, the combined trajectories reveal the geometry of the recording volume and what paths individuals prefer.

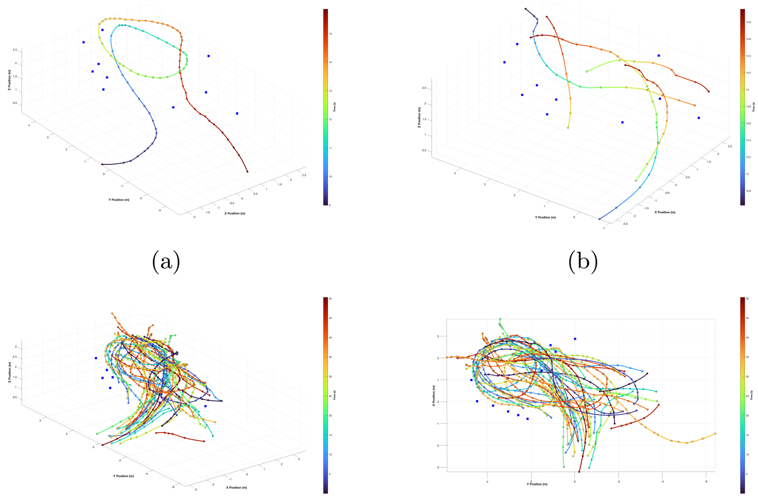
Table 2: Summary of cumulative time passed tracking bats categorized by number of present individuals.

Number of Bats	**Number of Instances**	**Cumulative Time (s)**
1 Bat	-	170.52
2 Bats	61	95.90
3 Bats	33	40.93
4 Bats	10	8.05
5 Bats	5	1.12

5.3 Comparing acoustic reconstruction to video reconstruction

The acoustically reconstructed trajectories are ground-truthed using independently and previously produced trajectories using video from three thermal cameras. In total, eight video trajectories were available for analysis, and of those, six had a corresponding acoustic trajectory. Because the two recording methods did not share a synchronized clock, the method chosen to compare the reconstructions measures the shortest distance from a reconstructed call position to the video trajectory that is orthogonal to the audio trajectory at the call position. This was done by projecting an orthogonal plane at every call position and finding the intersection of the plane and the video trajectory. The video trajectories and the corresponding audio trajectories for two separate recordings are provided in Figure 12. A total of 64 call positions were compared with a video reconstruction, with one distance being removed due to a large jitter in the video trajectory, resulting in N = 63 points available for evaluating the pipeline.

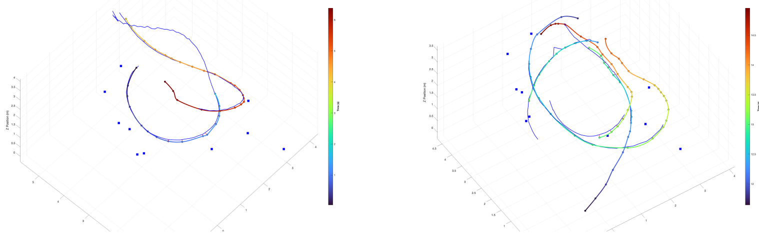
The error distances are visualized in a box plot in Figure 13. The range of the 3D errors were $0.05 - 0.10$ m, in the X-axis the range was $0.00 - 0.07$ m, in the Y-axis $0.00 - 0.09$ m, and in the Z-axis $0.00 - 0.08$ m. The results of the comparison are summarized in Table 3, which shows that the acoustic



(a) (b)

(c) (d)

Figure 11: Trajectories reconstructed from audio. The color gradient represents the time passed during flight. The dots in the trajectories are the estimated call positions, with a cubic Hermite interpolation connecting them. Each blue square represents a microphone. (a) displays the longest reconstructed trajectory, capturing 25.52 m and 6.87 s of flight. Visualized in (b) is one of the instances with the highest number of simultaneous trajectories. At one point 5 bats are successfully tracked at the same time. (c) and (d) show the all the 51 reconstructed trajectories from a 50.6 s long audio recording.



(a) (b)

Figure 12: Video trajectories overlaid with the corresponding audio trajectories. The video trajectories are colored a solid blue, while the audio trajectories are colored using a gradient representing the passage of time. Red lines between call positions and video trajectories visualize the orthogonal error distance. In (a) two matching trajectories are shown, and in (b) four matching trajectories are shown.

reconstruction from the proposed pipeline deviates from the video trajectory with a MAE of 0.036 m and an RMSE of 0.042 m. This result validates the utility of the pipeline. The small difference of 6 mm between MAE and RMSE highlights that the reconstruction is both highly accurate and consistent, with few large deviations from the ground truth. However, note that the error in the Y-axis has a markedly higher MAE and difference between RMSE and MAE than the X or Z-axes. The larger error could be due to the design of the microphone array influencing the quality of reconstructions in certain directions, or simply due to the size of the data set being analyzed. In order to compare this result to what de Framond at al. [15] achieved, the differences between the two methods is first made clear. This thesis defined the error as the absolute orthogonal distance between an estimated position and a ground truth position, whereas de Framond et al. measured the error as the difference in distance from central microphone to both an estimated position and a ground truth position. This means that de Framond et al. could have potentially reported an error of 0 m if the estimated position and ground truth position had the same radial distance from the central microphone, but different coordinates. This error also depended on the distance from the central microphone, something not explicitly measured for the purposes of this thesis. However, as previously shown, de Framond et al. produced position estimates with an MAE of around 0.1 m at 3 m away from the central microphone, a distance roughly comparable to the operating distance between source and microphones in the Ushichka array (5.5 m across). Given that the proposed pipeline achieves an overall MAE of 0.036 m using a stricter error metric, this comparison strongly indicates that the accuracy of the proposed pipeline exceeds current state-of-the-art methods in the field.

To explain the possible sources of the error when comparing the acoustically reconstructed trajectories to those constructed from video, we begin by investigating the video reconstruction. Firstly, the video tracking follows single points on the body of a bat, which means that there is inherent uncertainty as to what specific part of the bat is tracked. For reference, the main species present in the Ushichka dataset, *Myotis myotis*, has an adult wingspan of around 40 cm [22]. Furthermore, acoustic tracking locates the source of the emitted sound, which in this case is the mouth of a bat, usually a few centimeters away from the center of the body. Secondly, there is a systematic error that arises when converting between the coordinate system of the video trajectory and the coordinate system defined by the acoustic microphone calibration outlined in Section 4.3. The size of this specific error is not verified but likely affects the comparison of the two reconstructions.

Table 3: Summary of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) between video and audio reconstructions, (N = 63).

Coordinate Axis	**MAE (m)**	**RMSE (m)**
Overall 3D error	0.036	0.042
X	0.015	0.021
Y	0.021	0.030
Z	0.016	0.022

There are also numerous sources of errors in the proposed pipeline to be discussed. To start off, the position of the microphones stemming from the calibration affects the geometry of the Multilateration problem. If we assume

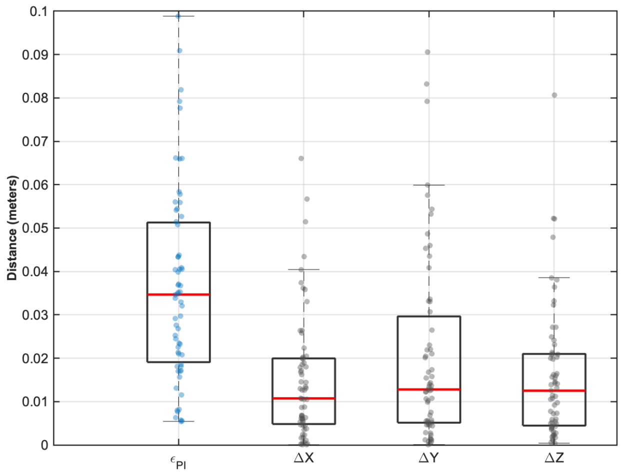


Figure 13: Box plots visualizing the spread in the orthogonal distance between audio and video reconstructions, (N = 63). Each dot represents a distance measurement. The left most box shows the total error distances ϵ_{PI}, with the remaining three showing the error distance in the X, Y, and Z direction.

error distances similar to those present in the office experiment performed by Batstone et al. [21], in the range of $5 - 20$ cm this might create a serious discrepancy between the estimated TDOA and the assumed geometry. This would be especially noticeable when the bat emits a call close to a microphone with a large positional error. However, this error is likely handled well by the RANSAC implementation, which serves the express purpose of mitigating the influence of individual outliers.

The automated call detection dictates which calls get sent for further analysis and therefore the quality of the detections matters in two ways: If a call never gets detected by BatNet it will never be used to reconstruct a trajectory, potentially leading to trajectories being cut short or separated into shorter segments. This problem is likely mitigated by including every recording channel in the detection pipeline, increasing the probability that a call will be detected in at least one channel. The other way the detection matters is in its accuracy. If a call is detected, but the exact alignment centers it slightly before the actual call or slightly after on a potential echo, this will affect the quality of the TDOA estimation depending on which parts of the call are present in the reference signal when handed to the GCC-PHAT estimation. To account for this, the reference window encapsulated 10 ms of audio, about double the length of a *Myotis myotis* call. Another challenge multiple parts of the pipeline faces, but is especially worth mentioning in relation to the automatic detection, is when multiple individuals are present simultaneously. Simultaneous call emissions can visually overlap in the spectrogram, or be close enough that BatNet considers one an echo of the other. The problem of correctly identifying calls during periods of high activity is not just challenging for trained CNNs, but for trained humans as well. This embodies an inherent limitation in how many sources that can be tracked simultaneously.

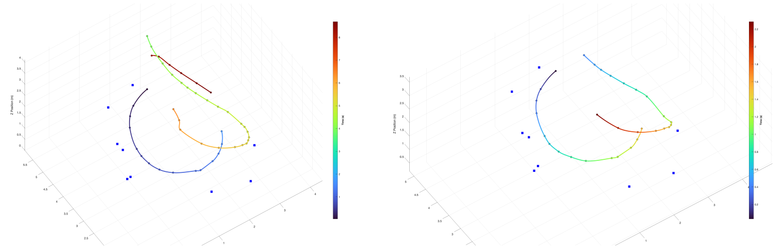
The TDOA estimation through GCC-PHAT is comparatively robust to reverberations, but even so, the estimations are still error prone. It is not rare that echoes are recorded with a higher amplitude than the actual call, sometimes leading to the Cross-Correlation score in turn being higher and an error in the estimation. Another source of errors concerning GCC-PHAT is when either the reference or target signal is weak or the signal to noise ratio is low. This can create multiple peaks in the Cross-Correlation score and the algorithm identifying the wrong TDOA. As long as at least four good TDOA estimations are made, RANSAC should be able to minimize these issues.

The Multilateration solver and optimizer coupled with RANSAC are very robust parts of the pipeline, and the overall effect is a strong mitigation of previous errors accumulating and severely affecting the final product. Worth discussing is the choice of a tolerance distance of 0.02 m corresponding to a maximum allowed difference between estimated and measured TDOA of $\approx 58.8\ \mu s$. This setting was chosen through experimentation and generally seems to strike the balance between being strict enough to be precise, and loose enough to allow for inliers. Solutions ranging from 4 inliers, which is the minimum, to 10 are all common and show that given a good detection and TDOA estimations, the minimal solver with RANSAC is able to find a solution fitting all the data. Aggregating the differences in solutions before and after optimizing in a recording with 4760 analyzed calls shows that the median distance shift in a solution is ≈ 6 mm, demonstrating the stability of the initial estimate of the solver. The average, however, is a shift of ≈ 173 m. This number arises due to a minority of solutions being extreme outliers, sometimes placed thousands of meters away from the array. These solutions likely arise from faulty call detections and TDOA estimates, and should not pose any issue for the final result, as they will simply not be appended to any trajectories.

The multi-hypothesis tracking and trajectory association strategy was developed ad hoc for the purpose of connecting the right call positions into the right trajectories. The method is the result of engineering trial and error, but it is verified by the results. The first step of consolidating points close in both space (< 15 cm) in time (< 10 ms) by selecting the one with the smallest RANSAC residual, greatly reduces the number of points used to construct hypotheses and mitigates the issue of duplicate trajectories being created from technically different but very similar groups of points. Since the multi-hypothesis tracker considers two trajectories as different if they don't share at least one point, clusters of points along a trajectory could exponentially increase the numbers of trajectory hypotheses being considered. This was not implementable using the available hardware. The physical limits for speed, acceleration and turn angle were verified to be reasonable by T. Beleyur through personal communication. However, the specific values representing realistic flight parameters have not been sufficiently studied in a cave environment. There is also the possibility that a bad solution that just happens to be placed favorably gets appended to a trajectory. If this point does not interfere with the rest of the trajectory this will currently go unnoticed, but if it instead leads a trajectory off course, this new hypothesis is unlikely to again append a random point, so this hypothesis will be discarded in favor of the one following the true trajectory. The case where random points probably do show up is in the very starts and ends of trajectories, where if a point is appendable, it wont interfere with the rest of the trajectory, and since it contains at least one more point than the true trajectory, it will be deemed as the most successful hypothesis. It is deemed very unlikely that any of the trajectories consist fully of random points, since this would mean that at least 6 random points were close enough in space and time, and aligned in a physically feasible way.

5.4 Test with Synthesized Audio

The reconstructed trajectories resulting from the test on synthesized audio data are presented in Figure 14. Of the original three trajectories, one was unable to be reconstructed from the synthesized audio, whereas two were mostly recovered. The fact that only two of the trajectories could be reconstructed possibly shows a limit of the number of simultaneous sources that the pipeline can reliable handle.



(a) (b)

Figure 14: Comparison between reconstructions of unaltered audio and synthesized multi-target audio. The color gradient represents the passage of time, and the points along the trajectories designate the call position. In (a) three reconstructed paths of non simultaneous bats are shown. In (b) are shown two paths from synthesized audio encapsulating the same events.

The reason why one path disappeared while the others mostly stayed intact is not entirely clear and could be the result of multiple different effects. It could stem from the automatic detection failing when there are too many sources present, potentially causing quieter signals to drown out. In the unaltered audio file, a total of 412 calls were detected across all ten channels, whereas only 282 calls were detected in the synthesized audio, indicating missed calls were a primary failure point. Another possible source of error is that GCC-PHAT might be more prone to give poor estimates for TDOAs when there are more signals present in an audio segment. This could affect the final position estimates of the RANSAC solver if too many TDOAs are faulty. However, for the two surviving trajectories, video ground-truth was available, and when calculating the orthogonal distance error between the acoustic and video reconstructions for the synthesized audio, similar results were obtained to those presented in the previous section. For $N = 29$ call positions reconstructed from synthesized audio, the MAE was 0.036 m and the RMSE was 0.043 m. This test represents a smaller sample size, but the results are almost identical, demonstrating that the quality of the produced reconstruction is highly accurate when multiple sources are present, even though parts of trajectories might be lost due to missed call detections.

6 Conclusions and Future Work

This thesis successfully demonstrated a robust, automated pipeline capable of reconstructing 3D flight trajectories of echolocating bats entirely from multi-channel audio recordings. By integrating computer vision, signal processing, and a multilateration solver coupled with a RANSAC framework, the system processed 553.1 s of audio to extract 335 continuous flight trajectories. When validated against ground-truth thermal video data (N = 63 calls), the acoustic reconstructions yielded an orthogonal Mean Absolute Error (MAE) of 3.6 cm. This result verifies the viability of the pipeline's approach, and, by extension, motivates the use of flexible, large microphone arrays to study bats non-invasively in their natural habitats.

In order to potentially improve the performance of the method, a few areas of possible future research are identified:

- Optimizing call detection: While the SqueezeNet Convolutional Neural Network (CNN) provided satisfactory results as a baseline implementation, alternative or custom-built architectures could be explored. Furthermore, increasing the volume of high-quality training data would likely improve precision and recall. However, future models must be tailored to the specific properties of the recording environment, the target bat species, and the overall recording quality. Completely different methodologies, such as matched filter-based detection, are also worth investigating.
- Microphone array design: Using additional microphones more strategically placed out in a recording volume could potentially capture a higher number of chirps, as well as facilitate more stable solutions of the minimal multilateration problem. In addition, a larger number of sensors could increase consensus, even with a substantial number of faulty measurements, allowing RANSAC to find the best solution more reliably.
- Trajectory association: The method for associating call positions into trajectories used physical flight parameters to build feasible flight paths, and thus, obtaining a deeper understanding of the behavior of bats in different environments could improve the precision of physics-based filters.

Ultimately, this thesis demonstrates that the inherent challenges in acoustic tracking, such as multi-path interference, noise, and multiple sources, can be successfully overcome using established tools and methods. The centimeter-level accuracy achieved by the pipeline establishes a strong methodological foundation not only for studying echolocating bats, but potentially for broader applications in acoustic localization and multi-target tracking.

References

[1] T. Beleyur, "Theoretical and empirical investigations of echolocation in bat groups," pp. 85–108, 2020. Chapter 4: The Ushichka dataset: a multi-channel, multi-sensor dataset to unravel active sensing in groups.

[2] M. Larsson, V. Larsson, K. Astrom, and M. Oskarsson, "Optimal trilateration is an eigenvalue problem," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5586–5590, IEEE, 2019.

[3] J. Smith and J. Abel, "The spherical interpolation method of source localization," *IEEE Journal of Oceanic Engineering*, vol. 12, no. 1, pp. 246–252, 1987.

[4] D. J. Torrieri, "Statistical theory of passive location systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-20, no. 2, pp. 183–198, 1984.

[5] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1526–1540, 2004.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[8] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6, IEEE, 2015.

[9] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135, IEEE, 2017.

[10] C. Eutizi and F. Benedetto, "On the performance improvements of deep learning methods for audio event detection and classification," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 141–145, IEEE, 2021.

[11] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[12] G. W. Pierce and D. R. Griffin, "Experimental determination of supersonic notes emitted by bats," *Journal of Mammalogy*, vol. 19, no. 4, pp. 454–455, 1938.

[13] K. D. Roeder, "Acoustic sensitivity of the noctuid tympanic organ and its range for the cries of bats," *Journal of insect physiology*, vol. 12, no. 7, pp. 843–859, 1966.

[14] J. C. Koblitz, "Arrayvolution: using microphone arrays to study bats in the field," *Canadian Journal of Zoology*, vol. 96, no. 9, pp. 933–938, 2018.

[15] L. de Framond, T. Beleyur, D. Lewanzik, and H. R. Goerlitz, "Calibrated microphone array recordings reveal that a gleaning bat emits low-intensity echolocation calls even in open-space habitat," *Journal of Experimental Biology*, vol. 226, no. 18, p. jeb245801, 2023.

[16] E. Verreycken, R. Simon, B. Quirk-Royal, W. Daems, J. Barber, and J. Steckel, "Bio-acoustic tracking and localization using heterogeneous, scalable microphone arrays," *Communications biology*, vol. 4, no. 1, p. 1275, 2021.

[17] E. Verreycken, W. Daems, and J. Steckel, "Passive acoustic sound source tracking in 3d using distributed microphone arrays," in *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–8, 2018.

[18] O. Mac Aodha, R. Gibb, K. E. Barlow, E. Browning, M. Firman, R. Freeman, B. Harder, L. Kinsey, G. R. Mead, S. E. Newson, *et al.*, "Bat detective—deep learning tools for bat acoustic signal detection," *PLoS computational biology*, vol. 14, no. 3, p. e1005995, 2018.

[19] S. Zhayida, F. Andersson, Y. Kuang, and K. Åström, "An automatic system for microphone self-localization using ambient sound," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 954–958, IEEE, 2014.

[20] Y. Kuang and K. Åström, "Stratified sensor network self-calibration from tdoa measurements," in *21st European Signal Processing Conference (EUSIPCO 2013)*, pp. 1–5, IEEE, 2013.

[21] K. Batstone, G. Flood, T. Beleyur, V. Larsson, H. Goerlitz, M. Oskarsson, and K. Åström, "Robust self-calibration of constant offset time-difference-of-arrival," in *2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019 - Proceedings*, pp. 4410–4414, IEEE - Institute of Electrical and Electronics Engineers Inc., 2019.

[22] R. Arlettaz, M. Ruedi, and J. Hausser, "Field morphological identifications of myotis myotis and myotis blythi (chiroptera, vespertilionidae): a multivariate approach," *Myotis*, vol. 29, pp. 7–16, 1991.

Master's Theses in Mathematical Sciences 2026:E56
ISSN 1404-6342

LUTFMA-3629-2026

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/