# Vocal Call Locator Benchmark (VCL) for localizing rodent vocalizations from multi-channel audio

**Ralph E Peterson**[1,2]*, **Aramis Tanelus**[2,*], **Christopher Ick**[3], **Bartul Mimica**[4], **Niegil Francis**[1,5], **Violet J Ivan**[1], **Aman Choudhri**[6], **Annegret L Falkner**[4], **Mala Murthy**[4], **David M Schneider**[1], **Dan H Sanes**[1], **Alex H Williams**[1,2†]

[1]NYU, Center for Neural Science
[2]Flatiron Institute, Center for Computational Neuroscience
[3]NYU, Center for Data Science
[4]Princeton Neuroscience Institute
[5]NYU, Tandon School of Engineering
[6]Columbia Univsersity

## Abstract

Understanding the behavioral and neural dynamics of social interactions is a goal of contemporary neuroscience. Many machine learning methods have emerged in recent years to make sense of complex video and neurophysiological data that result from these experiments. Less focus has been placed on understanding how animals process acoustic information, including social vocalizations. A critical step to bridge this gap is determining the senders and receivers of acoustic information in social interactions. While sound source localization (SSL) is a classic problem in signal processing, existing approaches are limited in their ability to localize animal-generated sounds in standard laboratory environments. Advances in deep learning methods for SSL are likely to help address these limitations, however there are currently no publicly available models, datasets, or benchmarks to systematically evaluate SSL algorithms in the domain of bioacoustics. Here, we present the VCL Benchmark: the first large-scale dataset for benchmarking SSL algorithms in rodents. We acquired synchronized video and multi-channel audio recordings of 767,295 sounds with annotated ground truth sources across 9 conditions. The dataset provides benchmarks which evaluate SSL performance on real data, simulated acoustic data, and a mixture of real and simulated data. We intend for this benchmark to facilitate knowledge transfer between the neuroscience and acoustic machine learning communities, which have had limited overlap.

**Data is available at:** `vclbenchmark.flatironinstitute.org`

## 1 Introduction

An ongoing renaissance of ethology in the field of neuroscience has shown the importance of conducting experiments in naturalistic contexts, particularly social interactions [39, 1]. Most experiments in social neuroscience have focused on relatively constrained contexts over short timescales, however an emerging paradigm shift is leading laboratories to adopt longitudinal experiments in semi-natural or natural environments [50]. With this shift comes significant data analytic challenges—such as how to track individuals in groups of socially interacting animals—necessitating collaboration between the fields of machine learning and neuroscience [10, 44].

*Equal Contribution
†Correspondence to `rep359@nyu.edu` and `alex.h.williams@nyu.edu`

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Substantial progress has been made in applying machine vision to multi-animal pose tracking and action recognition [45, 32, 37, 56], however applications of machine audio for acoustic analysis of animal generated social sounds (e.g. vocalizations or footstep sounds) have only recently begun [49, 20]. To study the dynamics of vocal communication and their neural basis, ethologists and neuroscientists have developed a multitude of approaches to attribute vocal calls to individual animals within an interacting social group, however many existing approaches for vocalization attribution necessitate specialized experimental apparatuses and paradigms that hinder the expression of natural social behaviors. For example, invasive surgical procedures, such as affixing custom-built miniature sensors to each animal [17, 48, 62], are often needed to obtain precise measurements of which individual is vocalizing. In addition to being labor intensive and species specific, these surgeries are often not tractable in very small or young animals, may alter an animal's natural behavioral repertoire, and are not scalable to large social groups. Thus, there is considerable interest in developing non-invasive sound vocal call attribution methods that work off-the-shelf in laboratory settings.

Sound source localization (SSL) is a decades old problem in acoustical signal processing, and several neuroscience groups have adapted classical algorithms from this literature to localize animal sounds [41, 54, 64]. These approaches can work reasonably well in specialized acoustically transparent environments, however they tend to fail in reverberant environments (see Supplement) that are required for next-generation naturalistic experiments.

Data-driven modeling approaches with fewer idealized assumptions may be expected to achieve greater performance [67]. Indeed, in the broader audio machine learning community, deep networks are commonly used to localize sounds [22]. Typically, these approaches have been targeted at human-scale acoustic environments—e.g. localizing sounds within rooms of a home [52]. To our knowledge, no benchmark datasets or deep network models have been developed for localizing sounds emitted by small animals (e.g. rodents) interacting in common laboratory environments (e.g. a spatial footprint less than one square meter). To address this, we present benchmark datasets for training and evaluating SSL techniques in reverberant conditions.

## 2 Background and Related Work

### 2.1 Existing Benchmarks

Acoustic engineers are interested in SSL algorithms for a variety of downstream applications. For example, localization can enable audio source separation [35] by disentangling simultaneous sounds emanating from different locations. Other applications include the development of smart home and assisted living technologies [19], teleconferencing [63], and human-robot interactions [33]. To facilitate these aims, several benchmark datasets have been developed in recent years including the L3DAS challenges [23, 24, 21], LOCATA challenge [15], and STARSS23 [52].

Notably, all of these applications and associated benchmarks are (a) focused on a range of sound frequencies that are human audible, and (b) focused on large environments such as offices and household rooms with relatively low reverberation. Our benchmark differs along both of these dimensions, which are important for neuroscience and ethology applications.

Many rodents vocalize and detect sounds in both sonic and ultrasonic ranges. For example, mice, rats, and gerbils collectively have hearing sensitivity that spans ~50-100,000 Hz with vocalizations spanning ~100-100,000 Hz [43]. Localizing sounds across a broad spectrum of frequencies introduces interesting complications to the SSL problem. Phase differences across microphones carry less reliable information for higher frequency sounds (see e.g. [28]). Moreover, a microphone's spatial sensitivity profile will generally be frequency dependent (see microphone specifications for ultrasonic condenser microphone CM16-CMPA from Avisoft Bioacoustics). Therefore, sounds emanating from the same location with the same source volume but distinct frequencies can register with unique level difference profiles across microphones. Thus, different acoustical computations are required to perform SSL for high and low frequency sounds. Indeed, we find that deep networks trained on low frequency sounds in our benchmark fail to generalize when tested on high frequency sounds (see Supplementary Figure 1).

Moreover, many model organisms (rodents, birds, and bats) are experimentally monitored in laboratory environments made of rigid and reverberant materials. The use of these materials is necessary to prevent animals from escaping experimental arenas, which is of particular concern when doing

longitudinal semi-natural experiments. For example, in attempts to mitigate reverberance using specialized equipment such as anechoic foam and acoustically transparent mesh, we found that gerbils will climb or chew through material after a short time in the arena. Therefore, use of hard plastic materials, even at the expense of being more reverberant, is required. Thus, the prevalence and character of sound reflections is a unique feature of the VCL benchmark. For variety, we also include benchmark data from an environment with sound absorbent wall material (E3).

### 2.2 Classical work on SSL in engineering and neuroscience

Conventional methods for SSL from acoustic signal processing are summarized in [12]. These methods primarily use differences in arrival times or signal phase across microphones to estimate sources; differences in volume levels are often ignored as a source of information (but see [3]). We use the Mouse Ultrasonic Source Estimation (MUSE) tool [41, 64] as a representative stand-in for these classic approaches in our benchmark experiments. An alternative method based on arrival times was recently proposed by Sterling, Teunisse, and Englitz [55] (see also [42]).

Neural circuit mechanisms of SSL have been extensively studied in model organisms like barn owls, which utilize exquisite SSL capabilities to hunt prey [30]. Neurons in the early auditory system represent both interaural timing and level differences in multiple animal species [9, 4, 7]. Behavioral studies in humans also establish the importance of both interaural timing and level differences [5], and the relative importance of these cues depends on sound frequency and the level of sound reverberation, among other factors [34, 28, 16]. Altogether, the neuroscience and psychophysics literature establishes that animals are adept at localizing sounds in reverberant environments. Moreover, in contrast to many classical SSL algorithms that leverage phase differences across audio waveforms, humans and animals use a complex combination of acoustical cues to localize sounds.

### 2.3 Deep learning approaches to SSL

SSL algorithms account for a variety of event-specific factors including sound frequency, volume, and reverberation. It is challenging to rationally engineer an algorithm to account for all of these factors and the acoustic machine learning community has therefore increasingly turned to deep neural networks (DNNs) to perform SSL. Grumiaux et al. [22] provide a recent and comprehensive review of this literature, including popular architectures, datasets, and simulation methods. Existing approaches to applying DNNs to SSL leverage a variety of input featurizations, like time-frequency representations (spectrograms) of the input audio. In our experiments, we use raw audio waveforms and DNNs with 1D convolutional layers, which are a reasonable standard for benchmarking purposes (see e.g. [61]). Similar to the existing SSL benchmarks listed above, the vast majority of published DNN models have focused on large home or office environments, which differ substantially from our applications of interest.

### 2.4 Acoustic simulations

Across a variety of machine learning tasks, DNNs tend to require large amounts of training data [26]. This is problematic, since it is labor intensive to collect ground truth localization data and curate the result to ensure accurate labels. To overcome this limitation, there is recent interest in leveraging acoustic simulations to supplement DNN training sets. Geometric acoustic simulations such as the image source method (ISM) [2] are popular, due to their relatively low computational cost, as well as their ability to preserve spatial information necessary for SSL [8][31]. Recent work has shown that use of room simulations generated using the ISM can also benefit model performance on real-world data [27] and can improve robustness by simulating a wider range of acoustic conditions than is present in an existing training dataset [47], despite perceptual limitations of the ISM. Given these trends in the field, our dataset release includes simulated environments and code for performing ISM simulations.

## 3 The VCL Dataset

The VCL Dataset consists of raw multi-channel audio and image data from 767,295 sound events with ground truth 2D position of the sound event source established by an overhead camera. We recorded synchronized audio (125 or 250 kHz sampling rate) and video (30 Hz or 150 Hz sampling rate) during

| Name | # Samples |
|---|---|
| **Speaker-4M-E1** | 70,914 |
| **Edison-4M-E1** | 266,877 |
| GerbilEarbud-4M-E1 | 7,698 |
| **SoloGerbil-4M-E1** | 61,513 |
| **DyadGerbil-4M-E1** | 653 |
| **Hexapod-8M-E2** | 156,900 |
| **MouseEarbud-24M-E3** | 200,000 |
| SoloMouse-24M-E3 | 549 |
| **DyadMouse-24M-E3** | 2,191 |

Table 1: Summary of datasets. Datasets in **blue** were used as training sets and for test sets when benchmarking SSL. Datasets in **red** were used as test sets when benchmarking sound attribution.

| Name | # Mics | Dimensions (m) |
|---|---|---|
| E1 | 4 | Top: 0.61595 x 0.41275<br>Bottom: 0.5588 x 0.3556<br>Height: 0.3683 |
| E2 | 8 | Top: 1.2182 x 0.9144<br>Bottom: 1.2182 x 0.9144<br>Height: 0.6096 |
| E3 | 24 | Top: 0.615 x 0.615<br>Bottom: 0.615 x 0.615<br>Height: 0.425 |

Table 2: Summary of environments. The final two characters in each dataset name (refer to Table 1) specifies the environment in which it was collected.

sound generating events from point sources emanating from either a speaker or real rodents. Sound events were sampled across three environments of varying size, microphone array geometries, and building material (Figure 1A-B, Table 1-2). Ground truth positions were extracted from the video stream using SLEAP [45] or OpenCV, and vocal events from real rodents were segmented from the audio stream using DAS [53]. To assess the quality of the machine-generated ground truth labels, we sampled 50 random vocal events from each training dataset and had four researchers manually label the ground truth location in each associated video frame (Supplementary Table 2). Timestamps from sound events using speaker playback were either recorded by a National Instruments data acquisition device or pre-computed and used to generate a wav file with known sound event onset times.

Brief descriptions of each dataset are included below and a more detailed description is provided in the supplemental datasheet (see "Collection Process" section). For datasets that involved speaker playback, we primarily used rodent vocalizations as stimuli (Figure 1C). In addition, we played sine sweeps in each environment which were used to compute a room impulse response (RIR, see Section 3.5). All procedures related to the maintenance and use animals were approved by the University Animal Welfare Committee at New York University and Princeton University. All experiments were performed in accordance with the relevant guidelines and regulations.

### 3.1 Speaker Datasets

The Speaker Dataset (Speaker-4M-E1) was generated by repeatedly presenting five characteristic gerbil vocal calls and a white noise stimulus at three volume levels (18 total stimulus classes) through an overheard Fountek NeoCd1.0 1.5" Ribbon Tweeter speaker. Between every set of presentations, the speaker was manually shifted two centimeters to trace a grid of roughly 400 points along the cage floor. This procedure yielded a dataset of 70,914 presentations spanning the 18 stimulus classes. Gerbil vocalizations can range in frequency from approximately 0.5-60 kHz and different vocalizations correspond to different types of social interactions in nature [58]. In this study, we selected a diverse set of commonly used vocal types vary in frequency range and ethologcial meaning.

### 3.2 Robot Datasets

The generation of the Speaker Dataset was quite labor intensive due to manual movement of the speaker, therefore the procedure was impractical for generating additional training datasets at numerical and spatial scale. To get around this issue, we developed two robotic approaches for autonomous playback of sound events. The Edison and Hexapod Datasets (Edison-4M-E1, Hexapod-8M-E2) were generated by periodically playing vocalizations through miniature speakers affixed to the robots as they performed a pseudo-random walk around the environment. The vocalizations used were sampled from a longitudinal recording of gerbil families [46].

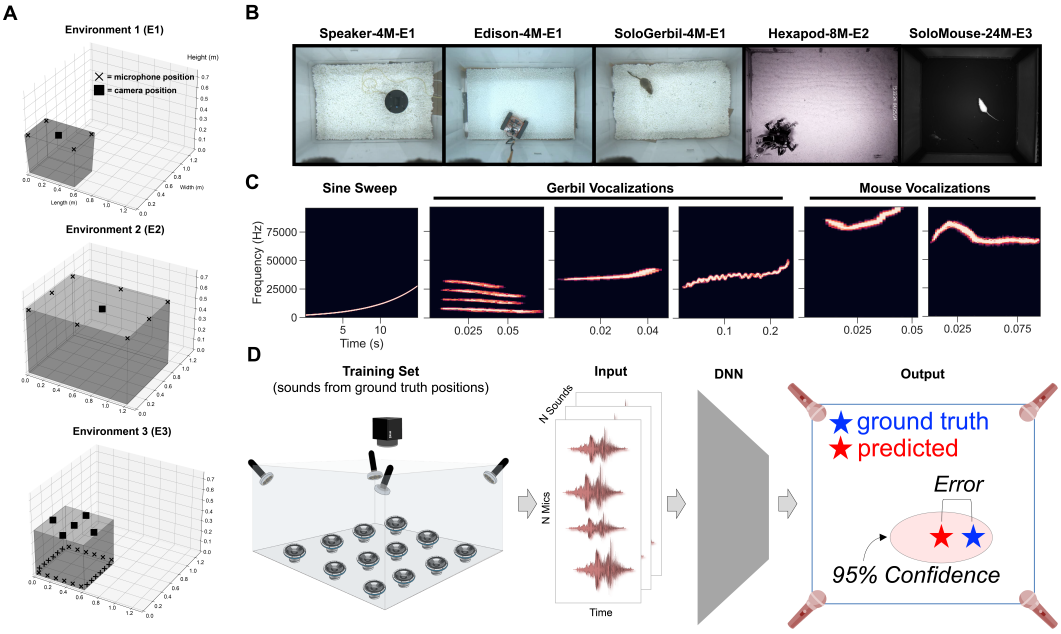


Figure 1: Overview of VCL benchmark. (A) Schematics of three laboratory arenas summarized in Table 2 showing relative size and positions of mics (X's) and cameras (squares). (B) Top-down views of different environments and training data generation modalities. (C) Examples of stimuli used for playback from Speaker, Edison, Earbud, and Hexapod datasets. (D) Schematic of pipeline depicting inputs (raw audio) and outputs (95% confidence interval).

### 3.3 Earbud Datasets

Speaker and robotic playback of vocalizations may not accurately represent the spatial usage and direction of vocalizations in real animals. To address this, we acquired two "Earbud" datasets (GerbilEarbud-4M-E1, MouseEarbud-24M-E3), in which gerbils or mice freely explored their environment with an earbud surgically affixed to their skull. We then played species typical vocalizations out of the earbud while animals exhibited a range of natural behaviors.

### 3.4 Solo/Dyad Gerbil & Mouse Datasets

Although isolated animals usually do not vocalize, we found that adolescent gerbils produce antiphonal responses to conspecific vocalizations played through a speaker. We leveraged this behavior to generate a large scale dataset, SoloGerbil-4M-E1, containing real gerbil-generated vocalizations in isolation. In addition, we elicited solo vocalizations in male mice (SoloMouse-24M-E3) by allowing female mice in estrus to explore the environment prior to male exploration.

Our ultimate goal is to use sound source estimates to attribute vocalizations to individuals in a group of socially interacting animals. To this end, we acquired vocalizations from pairs of interacting gerbils and mice (DyadGerbil-4M-E1, DyadMouse-24M-E3). Although we are unable to determine the ground truth position of vocalizations recorded from these interactions, we do know the locations of both potential sources and can therefore ascertain whether our model generates predictions with zero, one, or two animals within its confidence interval (See Task 2 below).

### 3.5 Synthetic Datasets

Since DNNs often require large training datasets and generation of datasets in the domain of SSL is laborious, we explored the use of acoustic simulations for supplementing real training data (Figure 2). We generated *in silico* models of our three environments accounting for physical measurements of the geometry, microphone placement, microphone directivity, and estimates of the material absorption coefficients (calculated via the inverse Sabine formula on room impulse response measurements with a sine sweep excitation). Code to reproduce these simulations and adapt them to new environments is

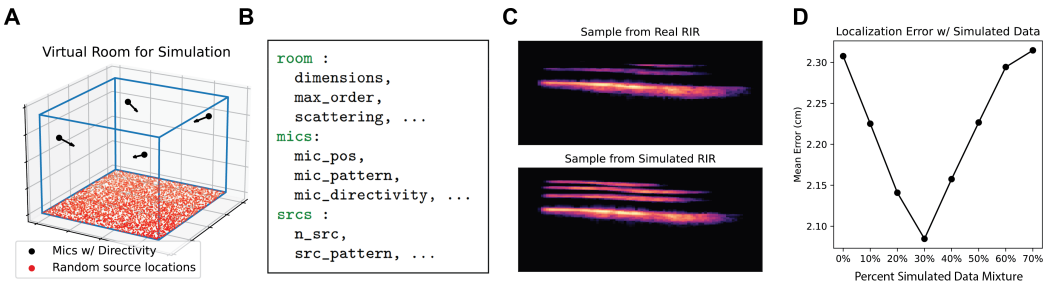


Figure 2: (A) Visualization of virtual room used for sythetic RIR generation via ISM (B) Sample of a room configuration YAML used to specify room geometry for simulations (C) Spectrograms comparing vocalizations convolved with recorded RIRs and simulated RIRs (D) Localization error as a function of added simulated data to the training corpus.

included in our code package accompanying the VCL benchmark. In preliminary experiments, we found that training DNNs on mixtures of real and simulated data can benefit performance (Figure 2D). DNNs trained exclusively on simulated data and evaluated on real data yields performance that marginally exceeds chance, but fails to match up to DNNs trained on smaller real datasets. This gap in performance indicates that our virtual acoustic models do not adequately simulate real acoustic environments. We believe that future work incorporating more robust acoustic simulations can bridge this gap. For these reasons, we do not include simulated data in benchmark experiments described below.

## 4 Benchmarks on VCL

We established a benchmark on the VCL Dataset using two distinct tasks.

- **Task 1 - Sound Source Localization:** Compare the performance of classical sound source localization algorithms with deep neural networks.
- **Task 2 - Vocalization Attribution:** Assign vocalizations to individuals in a dyad.

We evaluated performance on Task 1 using datasets with a single sound source (marked in **blue** in Table 1). We calculated the centimeter error between ground truth and predicted positions. Our aim is to achieve errors less than or equal to ~1 cm, as this is the approximate resolution required to attribute sound events to individual animals.[3] We also sought to benchmark the accuracy of model-derived confidence intervals. That is, for each prediction the model should produce a 2D set that contains the sound source with specified confidence (e.g. a 95% confidence set fail to contain the true sound source on only 5% of test set examples). Following procedures from Guo et al. [25], we plot reliability diagrams and report the expected calibration error (ECE) and maximal calibration error (MCE).

We evaluated performance on Task 2 using datasets with two potential sound sources (marked in **red** in Table 1). For Task 2, we report the number of animals inside the 95% confidence set of model predictions. For each sound event, the model can predict zero, one, or two animals within its confidence set. We report the frequency of each of these outcomes and interpret them as follows. First, if only one animal is within the confidence set, the model attributes the vocalization to that animal. We cannot for verify whether this attribution is correct because (unlike the datasets used in Task 1) we do not have ground truth measurements of the sound source. Second, if two animals are within the confidence set, then the model is unable to reliably attribute the sound to an individual. This outcome is neither correct nor incorrect. Finally, if zero animals are within the confidence set, then the model has falsely attributed the sound to a region. This outcome is clearly incorrect and should ideally happen less than 5% of the time when using a 95% confidence set.

[3] See, for example, Figure 1D in [55] for a distribution of inter-animal distances during natural social behavior.

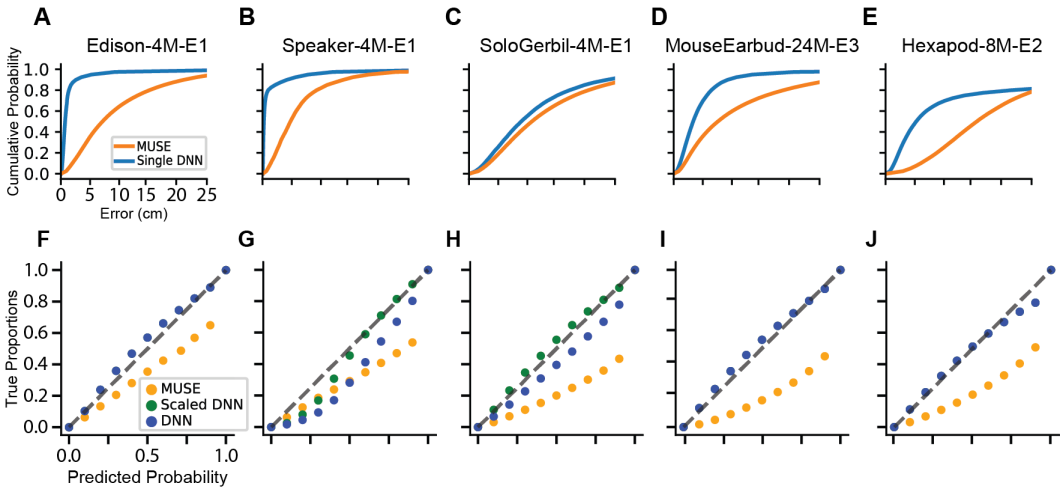


Figure 3: Benchmark performance. (A-E) Cumulative error distributions for MUSE and neural networks. (F-J) Reliability diagrams for MUSE (orange) and neural networks with (green) and without (blue) temperature scaling on heldout data from each dataset.

### 4.1 Convolutional Deep Neural Network

The network consists of 1D convolutional blocks connected in series. The network takes in raw multi-channel audio waveforms and outputs the mean and covariance of a 2D Gaussian distribution over the environment. Intuitively, the mean represents the network's best point estimate of the sound source and the scale and shape of the covariance matrix corresponds to an estimate of uncertainty. The network is trained with respect to labeled 2D sound source positions to minimize a negative log likelihood criterion—this is a proper scoring rule [18] which encourages the model to accurately portray its confidence in the predicted covariance. That is, the 95% upper level set of the Gaussian density should ideally act as a 95% confidence set. However, in line with previous reports, we sometimes observe that DNN confidence intervals are overconfident. In these cases, we use a temperature scaling procedure to calibrate the confidence intervals [25]. Further details on data preprocessing, model architecture, training procedure are provided in the Supplement.

### 4.2 MUSE Baseline Model

We compare the DNNs to a delay-and-sum beamforming approach used by neuroscientists called MUSE [41, 64]. MUSE works by computing cross-correlation signal between all pairs of microphone signals across hypothesized sound source locations, using the distance between microphones and the speed of sound to compute arrival time delays. The location that maximizes the summed response power over all microphones is then selected as a point estimate. We generate 95% confidence sets using a jackknife resampling technique proposed in Warren, Sangiamo, and Neunuebel [64].

### 4.3 Task 1 Results

Deep neural networks consistently produced estimates closer to the ground truth source than MUSE (Figure 3 A-E, Table 3). DNN performance was particularly strong on the Edison-4M-E1 and Speaker-4M-E1 datasets, achieving $<1$ cm error on 80.6% and 66.0% on the respective test sets. As mentioned above, this level of resolution should enable attribution of most vocalizations in realistic social encounters in rodents [55]. DNNs also outperformed MUSE on the remaining three datasets; however, they achieved sub-centimeter errors on less than 10% of the test set in all cases.

Moreover, we found that DNNs provide more accurate estimates of uncertainty relative to MUSE, as calculated by ECE and MCE (Table 4). This performance difference is visible in reliability diagrams, which show that MUSE predictions are over-confident (Figure 3F-J).

| Dataset | DNN Error (cm) | | | MUSE Error (cm) | | |
|---|---|---|---|---|---|---|
| | Mean | Median | % <1cm | Mean | Median | % <1cm |
| Speaker-4M-E1 | 1.4 | 0.2 | 80.6% | 6.4 | 4.8 | 5.9% |
| Edison-4M-E1 | 1.4 | 0.7 | 66.0% | 9.5 | 7.1 | 3.1% |
| SoloGerbil-4M-E1 | 12.0 | 10.0 | 1.0% | 13.2 | 10.8 | 1.0% |
| Hexapod-8M-E2 | 12.9 | 5.2 | 4.8% | 18.1 | 15.6 | 0.3% |
| MouseEarbud-24M-E3 | 4.1 | 2.6 | 8.7% | 11.3 | 7.6 | 3.3% |

Table 3: Summary of sound source localization errors for Task 1.

| Dataset | DNN | | Scaled DNN | | MUSE | |
|---|---|---|---|---|---|---|
| | ECE | MCE | ECE | MCE | ECE | MCE |
| Speaker-4M-E1 | 0.13 | 0.23 | 0.05 | 0.13 | 0.17 | 0.36 |
| Edison-4M-E1 | 0.03 | 0.07 | - | - | 0.12 | 0.25 |
| SoloGerbil-4M-E1 | 0.08 | 0.13 | 0.03 | 0.06 | 0.24 | 0.47 |
| Hexapod-8M-E2 | 0.03 | 0.11 | - | - | 0.22 | 0.40 |
| MouseEarbud-24M-E3 | 0.03 | 0.06 | - | - | 0.25 | 0.45 |

Table 4: Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) for Task 1.

### 4.4 Task 2 Results

To test the ability of our DNNs to assign vocalizations to individuals in dyadic interactions, we used DNNs trained on single-agent datasets, MouseEarbud-24M-E3 and SoloGerbil-4M-E1 respectively, to compute confidence bounds on vocalizations from the dyadic datasets MouseDyad-24M-E3 and GerbilDyad-4M-E1. As described above, we used temperature rescaling to ensure DNN confidence sets were well-calibrated. While we were capable of assigning between 19-29% of these calls to a single animal, over half of the vocalizations in each interaction yielded a confidence bound containing both animals (Table 5). Methods to resolve these shortcomings remain a focus of future work.

## 5 Limitations

Neuroscientists are interested in localizing sounds across a broad range of settings. We aimed to cover multiple rodent species (gerbils and mice), environment sizes, and microphone array geometries in this initial release. We also leveraged robots and head-mounted earbud speakers to collect sounds with known ground truth. However, this benchmark does not yet cover all use cases in neuroscience. Other commonly used model species—e.g., marmosets[14], bats[59], and various bird species[6]—are of great interest and are not covered by the current benchmark. Our experiments show that deep neural networks trained to localize sounds can fail to generalize across vocal call types (see Supplementary Figure 1). It would therefore be valuable to expand this benchmark to include a wider variety of animal species, call types, and increase the number of training samples. To this end, we include additional datasets which were not used in Task 1 due to their relatively small size (GerbilEarbud-4M-E1, SoloMouse-24M-E3), which will aid future experiments assessing generalization performance across datasets (e.g. train on Speaker-4M-E1, predict on GerbilEarbud-4M-E1).

Our current benchmark only provides images from a single camera view, which can be used to localize sounds in 2D. While this agrees with current practices within the field [41, 54, 38] and is in line with the equipment readily available to most labs, it is insufficient to infer 3D body pose information. One could imagine that knowing the 3D position and 3D heading direction of a vocalizing rodent could provide a more rich and effective supervision signal to train a deep network. A number of 3D pose tracking tools for animal models have been developed in very recent years [66, 40, 29, 36, 13]. These tools could be leveraged if future benchmarks collect multiple camera views. Ultimately, it would be

| | Gerbil Dyad | | | Mouse Dyad | | |
|---|---|---|---|---|---|---|
| # Animals Captured | 0 | 1 | 2 | 0 | 1 | 2 |
| Percentage | 6.1% | 28.6% | 65.2% | 8.9% | 19.4% | 71.7% |

Table 5: Vocalization attribution results. Number of animals captured within the 95% confidence set.

useful to compare performance across 3D and 2D benchmarks, to ascertain whether the sound source localization problem is indeed easier in one setting or the other.

## 6 Discussion

SSL is a well-known and challenging problem. We collected a variety of datasets and developed benchmarks to assess these challenges in the context of neuroethological experiments in vocalizing rodents. This involves localizing sounds in reverberant environments across a very broad frequency range (including ultrasonic events), distinguishing our work from more standard SSL benchmarks and algorithms. Our experiments reveal that DNNs are a promising approach. In controlled settings (Edison-4M-E1 and Speaker-4M-E1 datasets), DNNs achieved sub-centimeter resolution. In larger environments (Hexapod-8M-E2) and in datasets with uncontrolled 3D variation in sound emissions (SoloGerbil-4M-E1 and MouseEarbud-24M-E3), DNN performance was less impressive, but still outperformed a well-established benchmark algorithm (MUSE), that is currently utilized.

In addition to continuing to experiment with advances in machine vision/audio, we are also interested in exploring performance improvements due to hardware optimization. Parameters such as number of microphones, their positions/directivity, and environment reverberance can all affect SSL performance. Future experiments will leverage acoustic simulations to explore this parameter space. Initial results suggest that varying the amount of reverberation in an environment drastically affects SSL performance and that this effect is more pronounced in MUSE than DNNs (see Supplementary Figure 3). Moreover, we assessed whether specific acoustic or environmental features within the dataset affect model performance (Supplementary Figure 4). Sound power and distance from center of environment have a compelling effect on performance, where low power sounds and sounds that occur far away from the center of the arena (i.e. close to the walls) are difficult to localize. Fundamental frequency does not have a strong relationship to performance.

The ultimate goal of most neuroscientists in this context is to attribute vocal calls to individuals amongst an interacting social group. Accurate SSL would enable this, but it is also possible to reframe this problem as a direct prediction task. Specifically, given a video and audio recording of $K$ interacting animals with ground truth labels for the source of each sound event, DNNs could be trained to perform $K$-way classification to identify the source. Future work should investigate this promising alternative approach, as it would enable DNNs to jointly leverage information from audio and video data as network inputs. On the other hand, we note several challenges that must be overcome. First, establishing ground truth in multi-animal recordings is non-trivial, though feasible in certain experiments [17, 48, 62]. Second, DNNs trained to process raw video can have trouble generalizing across recording sessions due to subtle changes in lighting or animal appearance [65, 51]. Finally, we note that at least $K = 2$ animals are required to make the problem nontrivial (when $K = 1$ the DNN could ignore the audio input to predict the source). It will be important to establish a flexible DNN architecture that can make accurate predictions even when the animal group size, $K$, is altered (see e.g. [68]). It is already possible to use the VCL datasets to explore these possibilities. For example, one could use audio and video data taken from the same or different sound events to train a DNN with a multimodal contrastive learning objective (see e.g. [57], for a related concept).

In summary, there are many promising, but under-investigated, machine learning methodologies for annotating vocal communication in rodents. The VCL benchmark is our attempt to spark a broader community effort to investigate the potential of these computational approaches. Indeed, collecting and curating these datasets is labor-intensive and in our case involved collaboration across multiple neuroscience labs. To our knowledge, very little (if any) comparable data containing raw audio and video from many thousands of rodent vocal calls currently exists in the public domain. Thus, we expect the VCL benchmark will enable new avenues of research within computational neuroscience.

**Acknowledgements and Ethics Statement**

We do not foresee any negative societal impacts arising from this work. We thank Megan Kirchgessner (NYU), Robert Froemke (NYU), and Marcelo Magnasco (Rockefeller) for discussions and suggestions regarding SSL applications in neuroscience. This work was supported by the National Institutes of Health R34-DA059513 (AHW, DHS, DMS), National Institutes of Health R01-DC020279 (DHS), National Institutes of Health 1R01-DC018802 (DMS, REP), National Institutes of Health Training Program in Computational Neuroscience T90DA059110 (REP), New York Stem Cell Foun-

dation (DMS), CV Starr Fellowship (BM), EMBO Postdoctoral Fellowship (BM), National Science Foundation Award 1922658 (CI).

## References


[1] Ralph Adolphs. "Conceptual challenges and directions for social neuroscience". In: *Neuron* 65.6 (2010), pp. 752–767.

[2] Jont B. Allen and David A. Berkley. "Image method for efficiently simulating small-room acoustics". In: *The Journal of the Acoustical Society of America* 65.4 (Apr. 1979), pp. 943–950. ISSN: 0001-4966. DOI: 10.1121/1.382599. eprint: https://pubs.aip.org/asa/jasa/article-pdf/65/4/943/11426543/943_1_online.pdf. URL: https://doi.org/10.1121/1.382599.

[3] S. Argentieri, P. Danès, and P. Souères. "A survey on sound source localization in robotics: From binaural to array processing methods". In: *Computer Speech & Language* 34.1 (2015), pp. 87–112. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2015.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0885230815000236.

[4] Go Ashida and Catherine E Carr. "Sound localization: Jeffress and beyond". In: *Current opinion in neurobiology* 21.5 (2011), pp. 745–751.

[5] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[6] Michael S Brainard and Allison J Doupe. "Translating birdsong: songbirds as a model for basic and applied medical research". In: *Annual review of neuroscience* 36 (2013), pp. 489–517.

[7] Catherine E Carr and Jakob Christensen-Dalsgaard. "Sound localization strategies in three predators". In: *Brain Behavior and Evolution* 86.1 (2015), pp. 17–27.

[8] Soumitro Chakrabarty and Emanuel A. P. Habets. "Broadband doa estimation using convolutional neural networks trained with noise signals". In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, Oct. 2017. DOI: 10.1109/waspaa.2017.8170010. URL: http://dx.doi.org/10.1109/WASPAA.2017.8170010.

[9] Yale E Cohen and Eric I Knudsen. "Maps versus clusters: different representations of auditory space in the midbrain and forebrain". In: *Trends in neurosciences* 22.3 (1999), pp. 128–135.

[10] Sandeep Robert Datta et al. "Computational neuroethology: a call to action". In: *Neuron* 104.1 (2019), pp. 11–24.

[11] Yann N. Dauphin et al. "Language Modeling with Gated Convolutional Networks". In: *CoRR* abs/1612.08083 (2016). arXiv: 1612.08083. URL: http://arxiv.org/abs/1612.08083.

[12] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein. "Robust localization in reverberant rooms". In: *Microphone arrays: signal processing techniques and applications*. Springer, 2001, pp. 157–180.

[13] Timothy W Dunn et al. "Geometric deep learning enables 3D kinematic profiling across species and environments". In: *Nature methods* 18.5 (2021), pp. 564–573.

[14] Steven J Eliades and Cory T Miller. "Marmoset vocal communication: behavior and neurobiology". In: *Developmental neurobiology* 77.3 (2017), pp. 286–299.

[15] Christine Evers et al. "The LOCATA Challenge: Acoustic Source Localization and Tracking". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 1620–1643. DOI: 10.1109/TASLP.2020.2990485.

[16] Andrew Francl and Josh H McDermott. "Deep neural network models of sound localization reveal how perception is adapted to real-world environments". In: *Nature human behaviour* 6.1 (2022), pp. 111–133.

[17] Makoto Fukushima and Daniel Margoliash. "The effects of delayed auditory feedback revealed by bone conduction microphone in adult zebra finches". In: *Scientific Reports* 5.1 (2015), p. 8800.

[18] Tilmann Gneiting and Adrian E Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477 (2007), pp. 359–378.

[19] Stefan Goetze et al. "Acoustic monitoring and localization for social care". In: *Journal of Computing Science and Engineering* 6.1 (2012), pp. 40–50.

[20] Jack Goffinet et al. “Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires”. In: *Elife* 10 (2021), e67855.

[21] Riccardo F Gramaccioni et al. “L3DAS23: Learning 3D Audio Sources for Audio-Visual Extended Reality”. In: *IEEE Open Journal of Signal Processing* (2024).

[22] Pierre-Amaury Grumiaux et al. “A survey of sound source localization with deep learning methods”. In: *The Journal of the Acoustical Society of America* 152.1 (2022), pp. 107–151.

[23] Eric Guizzo et al. “L3DAS21 challenge: Machine learning for 3D audio signal processing”. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2021, pp. 1–6.

[24] Eric Guizzo et al. “L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment”. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 9186–9190. DOI: 10.1109/ICASSP43922.2022.9746872.

[25] Chuan Guo et al. “On calibration of modern neural networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[26] Joel Hestness et al. “Deep learning scaling is predictable, empirically”. In: *arXiv preprint arXiv:1712.00409* (2017).

[27] Christopher Ick and Brian McFee. *Leveraging Geometrical Acoustic Simulations of Spatial Room Impulse Responses for Improved Sound Event Detection and Localization*. 2023. arXiv: 2309.03337 [eess.AS].

[28] Antje Ihlefeld and Barbara G Shinn-Cunningham. “Effect of source spectrum on sound localization in an everyday reverberant room”. In: *The Journal of the Acoustical Society of America* 130.1 (2011), pp. 324–333.

[29] Pierre Karashchuk et al. “Anipose: A toolkit for robust markerless 3D pose estimation”. In: *Cell reports* 36.13 (2021).

[30] Eric I. Knudsen. “Instructed learning in the auditory localization pathway of the barn owl”. In: *Nature* 417.6886 (2002), pp. 322–328. ISSN: 1476-4687. DOI: 10.1038/417322a. URL: https://doi.org/10.1038/417322a.

[31] Daniel Krause, Archontis Politis, and Konrad Kowalczyk. “Data Diversity for Improving DNN-based Localization of Concurrent Sound Events”. In: *2021 29th European Signal Processing Conference (EUSIPCO)*. 2021, pp. 236–240. DOI: 10.23919/EUSIPCO54536.2021.9616284.

[32] Jessy Lauer et al. “Multi-animal pose estimation, identification and tracking with DeepLabCut”. In: *Nature Methods* 19.4 (2022), pp. 496–504.

[33] Xiaofei Li et al. “Reverberant sound localization with a robot head based on direct-path relative transfer function”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 2819–2826.

[34] Ewan A Macpherson and John C Middlebrooks. “Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited”. In: *The Journal of the Acoustical Society of America* 111.5 (2002), pp. 2219–2236.

[35] Shoji Makino. *Audio source separation*. Vol. 433. Springer, 2018.

[36] Jesse D Marshall et al. “Continuous whole-body 3D kinematic recordings across the rodent behavioral repertoire”. In: *Neuron* 109.3 (2021), pp. 420–437.

[37] Jesse D Marshall et al. “The PAIR-R24M Dataset for Multi-animal 3D Pose Estimation”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021. URL: https://openreview.net/forum?id=-wVVl_UPr8.

[38] Jumpei Matsumoto et al. “Acoustic camera system for measuring ultrasound communication in mice”. In: *Iscience* 25.8 (2022).

[39] Cory T Miller et al. “Natural behavior is the language of the brain”. In: *Current Biology* 32.10 (2022), R482–R493.

[40] Tanmay Nath et al. “Using DeepLabCut for 3D markerless pose estimation across species and behaviors”. In: *Nature protocols* 14.7 (2019), pp. 2152–2176.

[41] Joshua P Neunuebel et al. “Female mice ultrasonically interact with males during courtship displays”. In: *eLife* 4 (2015). Ed. by Peggy Mason, e06203. ISSN: 2050-084X. DOI: 10.7554/eLife.06203. URL: https://doi.org/10.7554/eLife.06203.

[42] Gabriel Oliveira-Stahl et al. “High-precision spatial analysis of mouse courtship vocalization behavior reveals sex and strain differences”. In: *Scientific Reports* 13.1 (2023), p. 5219.

[43] Gleich Otto and Strutz Jürgen. “The Mongolian gerbil as a model for the analysis of peripheral and central age-dependent hearing loss”. In: *Hearing Loss* (2012).

[44] Talmo D Pereira, Joshua W Shaevitz, and Mala Murthy. “Quantifying behavior to understand the brain”. In: *Nature neuroscience* 23.12 (2020), pp. 1537–1549.

[45] Talmo D Pereira et al. “SLEAP: A deep learning system for multi-animal pose tracking”. In: *Nature methods* 19.4 (2022), pp. 486–495.

[46] Ralph E Peterson et al. “Unsupervised discovery of family specific vocal usage in the Mongolian gerbil”. In: *eLife* (2023), e89892.1. DOI: 10.7554/eLife.89892.1. URL: https://doi.org/10.7554/eLife.89892.1.

[47] Iran R. Roman et al. *Spatial Scaper: A Library to Simulate and Augment Soundscapes for Sound Event Localization and Detection in Realistic Rooms*. 2024. arXiv: 2401.12238 [eess.AS].

[48] Maimon C. Rose et al. “Cortical representation of group social communication in bats”. In: *Science* 374.6566 (2021), eaba9584. DOI: 10.1126/science.aba9584. eprint: https://www.science.org/doi/pdf/10.1126/science.aba9584. URL: https://www.science.org/doi/abs/10.1126/science.aba9584.

[49] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. “Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires”. In: *PLoS computational biology* 16.10 (2020), e1008228.

[50] Yair Shemesh and Alon Chen. “A paradigm shift in translational psychiatry through rodent neuroethology”. In: *Molecular psychiatry* 28.3 (2023), pp. 993–1003.

[51] Changhao Shi et al. “Learning Disentangled Behavior Embeddings”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 22562–22573. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/be37ff14df68192d976f6ce76c6cbd15-Paper.pdf.

[52] Kazuki Shimada et al. “STARSS23: An Audio-Visual Dataset of Spatial Recordings of Real Scenes with Spatiotemporal Annotations of Sound Events”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 72931–72957. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/e6c9671ed3b3106b71cafda3ba225c1a-Paper-Datasets_and_Benchmarks.pdf.

[53] Elsa Steinfath et al. “Fast and accurate annotation of acoustic signals with deep neural networks”. In: *Elife* 10 (2021), e68837.

[54] Max L Sterling, Ruben Teunisse, and Bernhard Englitz. “Rodent ultrasonic vocal interaction resolved with millimeter precision using hybrid beamforming”. In: *eLife* 12 (2023). Ed. by Brice Bathellier, e86126. ISSN: 2050-084X. DOI: 10.7554/eLife.86126. URL: https://doi.org/10.7554/eLife.86126.

[55] Max L Sterling, Ruben Teunisse, and Bernhard Englitz. “Rodent ultrasonic vocal interaction resolved with millimeter precision using hybrid beamforming”. In: *Elife* 12 (2023), e86126.

[56] Jennifer J. Sun et al. “The Multi-Agent Behavior Dataset: Mouse Dyadic Social Interactions”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021. URL: https://openreview.net/forum?id=NevK78-K4bZ.

[57] Weixuan Sun et al. “Learning audio-visual source localization via false negative aware contrastive learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6420–6429.

[58] Maria Ter-Mikaelian, Wipula B Yapa, and Rudolf Rübsamen. “Vocal behavior of the Mongolian gerbil in a seminatural enclosure”. In: *Behaviour* 149.5 (2012), pp. 461–492.

[59] Nachum Ulanovsky and Cynthia F Moss. “What the bat’s voice tells the bat’s brain”. In: *Proceedings of the National Academy of Sciences* 105.25 (2008), pp. 8491–8498.

[60] Aäron van den Oord et al. “WaveNet: A Generative Model for Raw Audio”. In: *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*. 2016, p. 125.

[61] Juan Manuel Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. “Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates”. In: *Sensors* 18.10 (2018), p. 3418.

[62] Elena N. Waidmann et al. "Mountable miniature microphones to identify and assign mouse ultrasonic vocalizations". In: *bioRxiv* (2024). DOI: 10.1101/2024.02.05.579003. eprint: https://www.biorxiv.org/content/early/2024/02/06/2024.02.05.579003.full.pdf. URL: https://www.biorxiv.org/content/early/2024/02/06/2024.02.05.579003.

[63] Hong Wang and Peter Chu. "Voice source localization for automatic camera pointing system in videoconferencing". In: *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 1997, pp. 187–190.

[64] Megan R Warren, Daniel T Sangiamo, and Joshua P Neunuebel. "High channel count microphone array accurately and precisely localizes ultrasonic signals from freely-moving mice". In: *Journal of neuroscience methods* 297 (2018), pp. 44–60.

[65] Matthew R. Whiteway et al. "Partitioning variability in animal behavioral videos using semi-supervised variational autoencoders". In: *PLOS Computational Biology* 17.9 (Sept. 2021), pp. 1–50. DOI: 10.1371/journal.pcbi.1009439. URL: https://doi.org/10.1371/journal.pcbi.1009439.

[66] Alexander B Wiltschko et al. "Mapping sub-second structure in mouse behavior". In: *Neuron* 88.6 (2015), pp. 1121–1135.

[67] Sean F Woodward, Diana Reiss, and Marcelo O Magnasco. "Learning to localize sounds in a highly reverberant environment: Machine-learning tracking of dolphin whistle-like sounds in a pool". In: *PloS one* 15.6 (2020), e0235155.

[68] Manzil Zaheer et al. "Deep Sets". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Acknowledgements and Ethics Statement.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See data website, "vocalator" GitHub repo for DNNs, and supplement.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplement.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Due to time constraints and anecdotal evidence that models from different random seeds produce similar results, we did not include error bars in this draft. We are happy to include them upon revision.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

## Supplementary Information

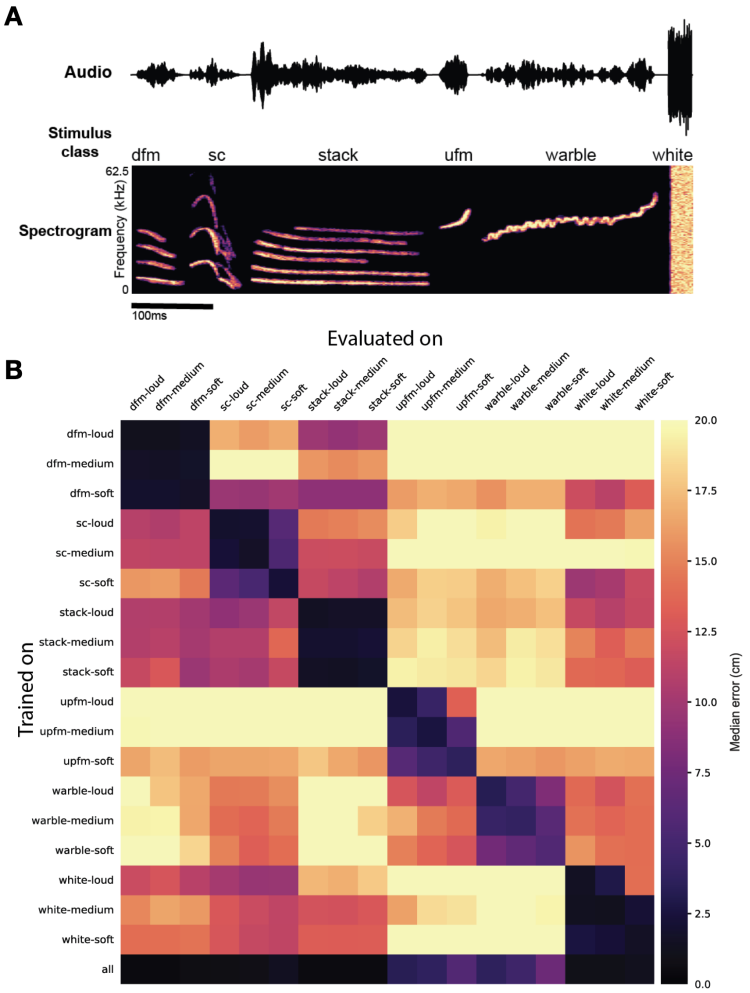


Figure 1: Generalizability across stimulus types. A.) Performance of models trained on single stimuli from Speaker-4M-E1 dataset and evaluated on all other stimulus types. (B) Stimuli used for speaker data set (dfm = down frequency modulated, sc = soft chirp, stack = harmonic stack, ufm = up frequency modulated)

Ultimately, we aim to create a tool that can be easily adapted by other labs which may have different recording environments. Additionally, we wish to utilize the tool for long-term recordings in which the types of vocalizations encountered may change over time as the animals enter new stages of life. As such, we have significant interest in the model's ability to generalize to unfamiliar vocal calls

To explore this, we tested the ability of deep networks to generalize to new vocal calls with different acoustic features. We partitioned the Speaker-4M-E1 Dataset according to stimulus type (Supplementary Figure 2A), trained a deep neural network on each subset, and measured its performance on every stimulus type individually (Supplementary Figure 2B). We found that while many models could generalize to new stimuli with performance exceeding chance, their ability to do so is greatly overshadowed by their performance on their own subsets. Models trained on a single stimulus type generalized well to the same stimulus at different volumes. (Supplementary Figure 2B, 3x3 block structure). This suggests that the networks are adapted to the statistics of the training set, and that training on a range of vocalizations with diverse spectral features will be necessary to achieve good performance across experimental cohorts, each of which may utilize slightly different vocal calls.

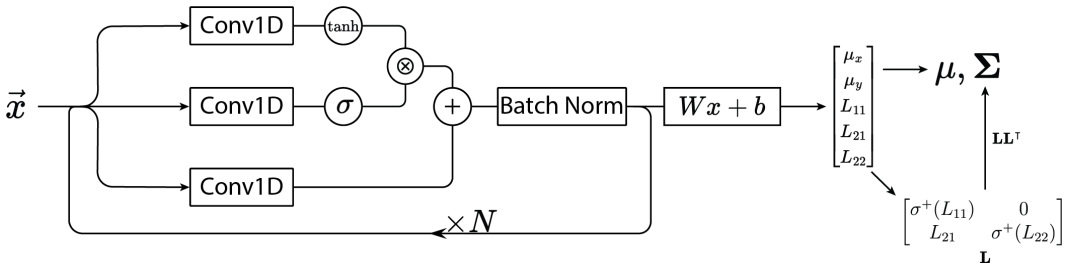


Figure 2: Network architecture.

| Layer | Channels | Downsample |
|---|---|---|
| 1 | 32 | No |
| 2 | 32 | Yes |
| 3 | 64 | No |
| 4 | 64 | Yes |
| 5 | 128 | No |
| 6 | 128 | Yes |
| 7 | 256 | No |
| 8 | 256 | Yes |
| 9 | 512 | No |
| 10 | 512 | Yes |

Table 1: Model Architecture Hyperparameters. Our model consists of 10 convolutional blocks. All use a kernel size of 33, dilation of 1, and stride of 1.

Mirroring gated linear units [11] and WaveNet [60], we apply tanh and sigmoid nonlinearities to the output of convolutions and multiply them element-wise. We add this product to the result of a third convolution and apply batch normalization to the sum. On layers with temporal downsampling, we perform average pooling with a stride and kernel size of 2 prior to normalization. On our datasets with four microphones, we incorporate pairwise cross-correlations of the microphone signals by concatenating the central elements of each cross-correlogram into a vector, passing it through a shallow MLP, and concatenating the result to the output of the final convolutional block. The model outputs the mean and covariance of a 2D Gaussian distribution with covariance specified by a Cholesky factor matrix. To parametrize the 2D gaussian posterior distribution, we first average the output of the final convolutional block over its time dimension and linearly project it to five components. Two of these determine the distribution's mean and the other three parametrize the Cholesky decomposition of the distribution's covariance matrix. In order to ensure the Cholesky factor has positive diagonals, we apply the softplus nonlinearity to the diagonal elements. During training, we evaluate the log likelihood of the ground truth positions with respect to the 2D Gaussians output by the network. We minimize the negative log likelihood using stochastic gradient descent with momentum. Throughout 50 epochs, we anneal the learning rate to 0 using a cosine schedule. We do not use weight decay.

For data preprocessing, we normalize the audio by ensuring a zero mean and unit variance across all elements, rather than scaling each channel individually. This approach ensures amplitude differences between channels are preserved after normalization. Throughout training, we apply various augmentations to the audio to enhance sample efficiency and performance on the validation set. As vocalization lengths vary substantially, we randomly crop them to a standardized length of 8192 samples (65.5ms at 125kHz) to facilitate batched computations. Additional augmentations include temporal masking, the introduction of white noise, and phase inversion. With the exception of cropping, which is applied universally to all samples, each augmentation has a 50% chance of being applied to a given vocalization.

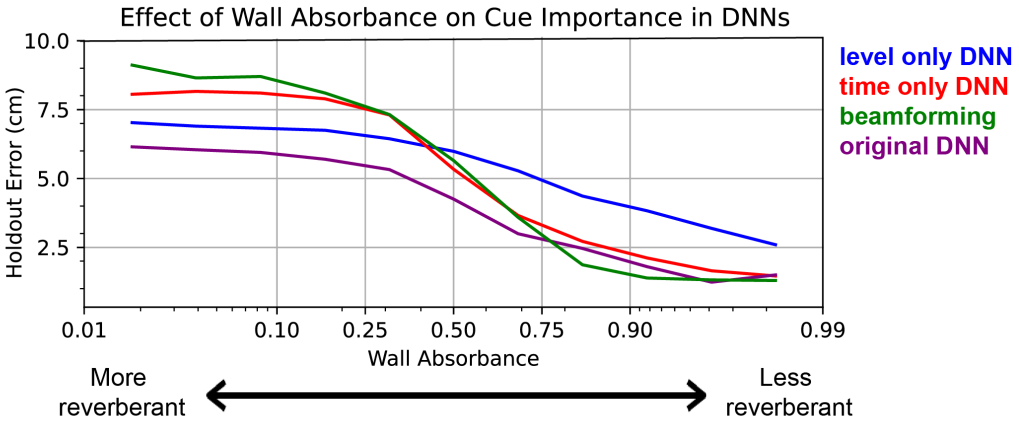


Figure 3: SSL performance with varying environmental reverberance.

We explored whether SSL performance systematically varied as a function of reverberance using acoustic simulations. First, we simulated an E1 environment, then simulated microphone signals from 50,000 gerbil vocalizations randomly sampled from [46]. Next, we compared DNN vs. MUSE (beamforming) performance and showed that DNNs (purple) outperform MUSE (green) in reverberant conditions and achieve equal performance in non-reverberant conditions. Furthermore, we explored which cues (temporal or level, i.e. akin to ITD and ILD cues used by animals) DNNs relied on for SSL. We created augmented training sets that either scrambled level differences between microphone channels (thereby only maintaining reliable time differences, red) or scrambled time differences (thereby only maintaining reliable level differences, blue). We find that time-only DNN performance matches MUSE, which is consistent with the fact that MUSE and other beamforming algorithms are time-only models. In addition, we find that level-only models outperform time-only models in reverberant conditions, but do worse in non-reverberant enrionments. Intriguingly, DNNs trained with both time and level (purple) perform better than level-only models in reverberant environments, suggesting that DNNs are making use of both available cues, though likely relying more on level. Future studies will aim to better understand how DNNs and biological neural networks balance the relative use of these two cues in reverberant listening conditions.

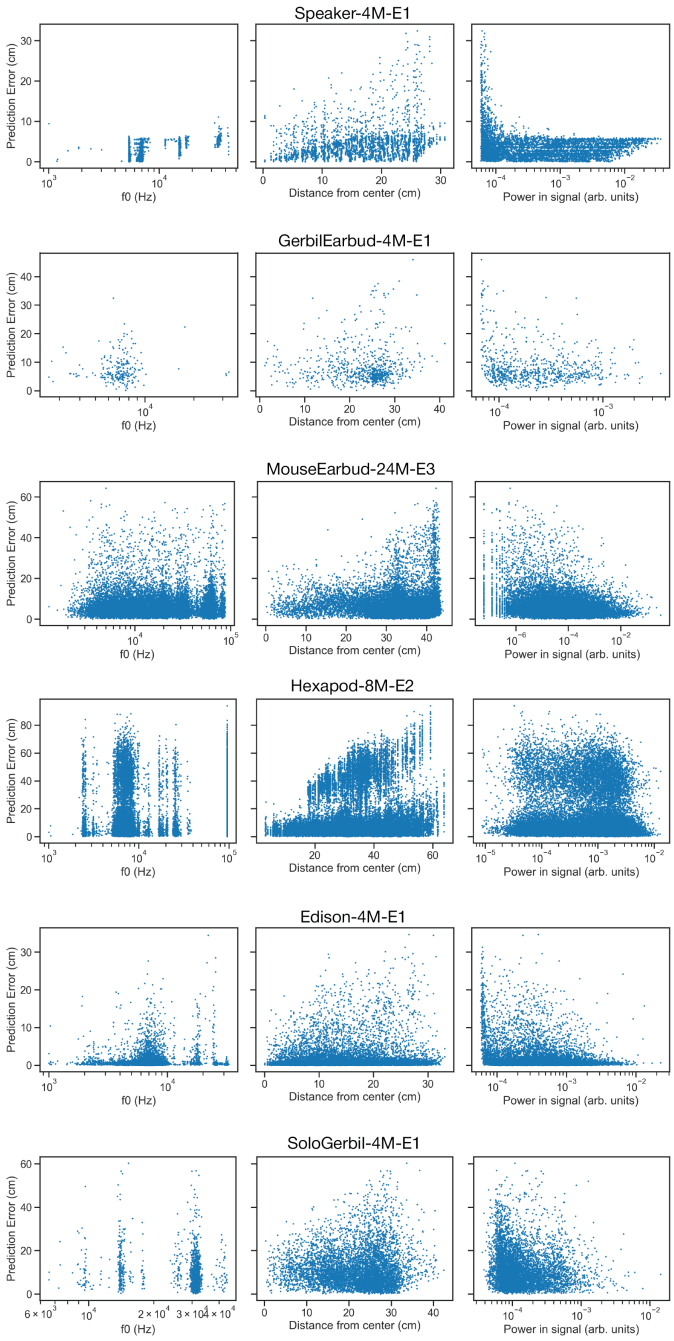


Figure 4: Effect of acoustic and environmental factors on localization performance.

To assess whether variation in localization performance relates to interpretable features in the dataset, we plotted the fundamental frequency, power, and distance to center of each sample in the test set as a function of localization error. Indeed, samples that are lower power and further from the center (i.e. next to the wall) are more difficult to localize. There is not an appreciable relationship between frequency of sample and localization error.

| Dataset | Error (px) | | | | |
|---|---|---|---|---|---|
| | Mean | Median | Max | Min | Human std |
| Speaker-4M-E1 | 5.8 | 6.2 | 12.3 | 0.8 | 0.6 |
| Edison-4M-E1 | 5.4 | 5.9 | 12.8 | 0.9 | 1.3 |
| SoloGerbil-4M-E1 | 28.8 | 18.5 | 184.6 | 7.0 | 1.0 |
| Hexapod-8M-E2 | 7.2 | 6.7 | 16.3 | 1.0 | 2.6 |
| MouseEarbud-24M-E3 | 4.9 | 4.2 | 23.9 | 0.7 | 1.5 |

Table 2: Analysis of error in machine-labeled ground truth.

Four researchers were tasked with annotating ground truth locations of the sound source within 50 video frames from each training dataset. We compared these human ground truth annotations with machine labeled ground truth locations used for SSL model training in this benchmark. The error in the machine label for each image was computed as the pixel distance between that label and the centroid of the human labels for the image in pixel space. We report the mean, median, maximum, and minimum error for each training dataset in addition to the average amount of deviation from the centroid in the human labels. SoloGerbil-4M-E1 exhibited a higher than expected error in machine-labeled ground truth locations, which at least partially explains the relatively high sound localization error for this dataset (Figure 3C, Table 3). Future releases of this benchmark will improve ground truth labels.