Benchmarks and deep learning models for localizing rodent vocalizations in social interactions

Ralph E. Peterson
NYU
Flatiron Institute

Aramis Tanelus
NYU
Flatiron Institute

Aman Chourdhri
Columbia University
Flatiron Institute

Violet Ivan
NYU

Aadi Prasad
Salk Institute
Flatiron Institute

David M. Schneider
NYU

Dan H. Sanes
NYU

Alex H. Williams
NYU
Flatiron Institute

Abstract

Social animals congregate in groups and vocalize to communicate. 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. Invasive surgical procedures, such as affixing custom-built miniature sensors to each animal, 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 and may alter an animal's natural behavioral repertoire. Thus, there is considerable interest in developing non-invasive sound source localization and vocal call attribution methods that work off-the-shelf in typical laboratory settings. To advance these aims in the domain of rodent neuroscience, we acquired synchronized video and multi-channel audio recordings with >300,000 annotated sound sources in small reverberant environments, and publicly release them as benchmarks. We then trained deep neural networks to localize and attribute vocal calls. This approach outperformed current protocols in the field, achieving ~5 mm accuracy on speaker-emitted sounds. Further, deep network ensembles produced well-calibrated estimates of uncertainty for each prediction. However, network performance was not robust to distributional shifts in the data, highlighting limitations and open challenges for future work.

1 Introduction

An ongoing renaissance of ethology in the field of neuroscience has shown the importance of conducting experiments in naturalistic contexts [1]. While this enables a more comprehensive understanding of the neurological underpinnings of behavior, it complicates analysis. For example many animals transmit social cues through vocalizations and other acoustic signals (e.g. foot thumping). Studying these in a naturalistic context requires knowing which animal within a freely interacting group is producing a sound. This is difficult to do with rodents because their vocalizations are not accompanied by a visually salient cue and thus must be localized aurally via an array of microphones or through invasive surgical procedures [2]. The resulting sound source localization (SSL) problem is decades old and many have adapted classical methods from signal processing to tackle it [2–4]. However, such methods assume no reverberations are present in the signal. In a complex environment designed to mimic an animal's natural habitat, this cannot be satisfied.

A data-driven modeling approach with fewer assumptions may be expected to overcome these challenges. Indeed, prior work has adapted deep networks to localize sounds in much larger environments—

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

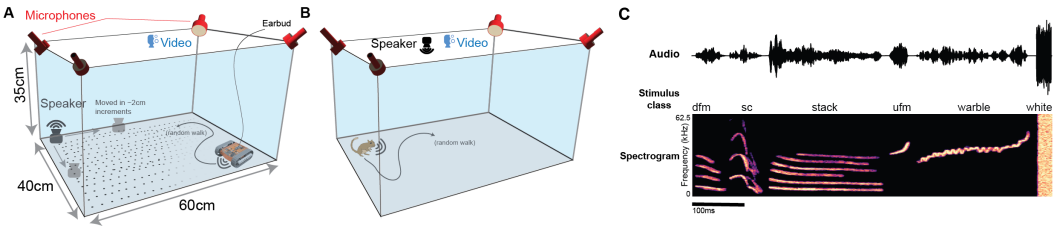

Figure 1: Experimental paradigm for generating training data. A.) An ultrasonic speaker hand-moved in 2 cm increments throughout the arena ("speaker" dataset). 6 classes of stimuli (see C) at 3 volumes levels were played out of the speaker at each position. Alternatively, a robot programmed to do a random walk with a wired earbud was used to play a related set of stimuli ("robot" dataset). Finally, adolescent gerbils respond vocally to playback of conspecific vocalizations. Therefore, we used a freely moving adolescent gerbil vocalizing to playback to generate a natural set of training data ("adolescent" dataset). C.) Stimuli used for speaker data set (dfm = down frequency modulated, sc = soft chirp, stack = harmonic stack, ufm = up frequency modulated).

e.g. human voices across multi-room home environments [5]. 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 more limited environments (e.g. a spatial footprint less than one square meter). To address this, we present three benchmark datasets for training and evaluating SSL techniques in reverberant conditions. In addition, we characterize the performance of ensembles of deep convolutional neural networks on SSL across these benchmarks.

2 Benchmark Datasets

We collected three benchmark datasets for SSL of rodent vocalizations, hereafter referred to as the Speaker, Robot, and Adolescent datasets. We have released the Speaker and Robot datasets publicly at: `https://github.com/neurostatslab/gerbilizer`. All datasets were collected in a rigid plastic cage (length = 60cm, depth = 40cm, height = 35cm) filled with ~1 inch of bedding. The ceiling of the cage was coated with acoustic foam. A FLIR Blackfly S USB3 camera and 4 Avisoft CM16/CMPA48AAF-5V microphones were used to acquire synchronized audio and video at 125kHz and 30fps respectively. Ground truth positions were extracted from the video stream through SLEAP [6] and vocal events were segmented from the audio stream using DAS [7]. Before use, all datasets were randomly split in an 8:1:1 ratio to generate training, validation, and testing sets respectively.

The Speaker Dataset was generated by repeatedly presenting five characteristic gerbil vocal calls and a white noise stimulus at three volume levels (18 total stimulus classes) through a Fountek NeoCd1.0 1.5" Ribbon Tweeter speaker inside the cage. 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. In this study, we selected a diverse set of commonly used vocal types vary in frequency range and ethologcial meaning.

The Robot Dataset was generated by periodically playing vocalizations through an earbud affixed to a Edison Robot V2.0 programmed to autonomously explore the arena. The vocalizations used were sampled from a longitudinal recording of gerbil families [8]. This yielded a much larger dataset of 287,792 presentations.

Finally, the Adolescent Dataset was generated by recording an adolescent gerbil's natural vocalizations. Although isolated animals often do not vocalize (see e.g. [9]), we found that adolescent gerbils produce antiphonal responses to a conspecific's vocalizations played through a speaker. The gerbil had no surgical implants or other abnormal conditions imposed on it and roamed the cage freely. This yielded a dataset of 22,371 vocalizations.

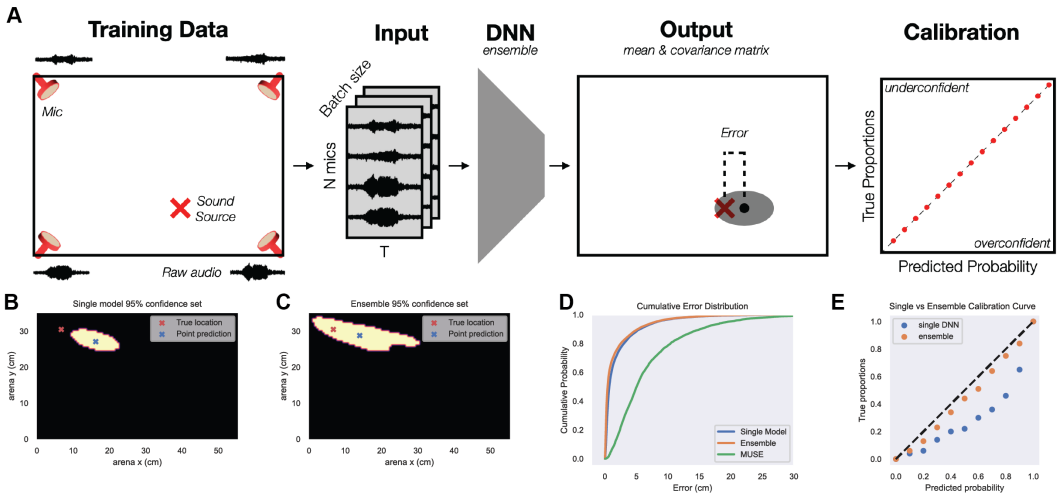

Figure 2: Deep learning pipeline for sound-source localization of rodent vocalizations. A.) Schematic of pipeline depicting inputs (raw audio) and outputs (calibrated 95% confidence set). See Sound Localization Results section for architecture details. B.) Example 95% confidence set for a single model. C.) Example 95% confidence set for an ensemble of 10 models. D.) Error CDF for single model, ensemble model, and MUSE (delay-and-sum beamforming). E.) Calibration curve for single vs. ensemble model.

3 Sound Source Localization Results

Equipped with these three datasets containing over 300,000 ground truth sound source annotations, we investigated the ability of deep networks to localize and attribute vocal calls. Our architecture consists of a sequence of convolutional blocks. As in WaveNet each block consists of an elementwise multiplication between the results of a filtering convolution and a gating convolution [10]. Unlike WaveNet, we forgo residual connections connecting all layers in order to allow some blocks to downsample the audio and reduce the computational cost of the model. We also use undilated, non-causal convolutions to maximize the model's ability to detect interaural time delay-like (ITD) cues that might be missed by coarser filters. In this vein, one of our future directions is to use equivariant convolutions over the affine group to better account for such effects that may occur at multiple time scales [11]. After 7 convolutional blocks, we take the mean of each channel across the time dimension and linearly project it to 5 values which parametrize the mean and covariance of the 2D Gaussian distribution (Fig. 2B). Finally, we minimize the negative log likelihood of the ground truth locations under this predictive Gaussian distribution. The mean and covariance can be respectively interpreted as the model's estimate of the sound source and its uncertainty about that estimate. Importantly, the negative log-likelihood criterion we use is a proper scoring rule, which is believed to encourage the model to produce well-calibrated (i.e. neither under- nor over-confident) estimates of uncertainty [12]. Additionally, we can combine the outputs of an ensemble of models by averaging these distributions to form a mixture of Gaussians (Fig. 2C).

As a baseline, we compare our model to MUSE, an existing method for localizing ultrasonic vocal sounds in anechoic environments. On our benchmark Speaker Dataset, our lone neural networks outperform MUSE (Fig. 2D), but are overconfident in their responses (Fig. 2E). Ensembling 10 of these models resulted in well-calibrated predictions as well as slightly increased accuracy (Fig. 2D-E). On our other datasets we similarly observed that neural networks provide more accurate predictions than MUSE and achieve better calibration when ensembled. The median error achieved by a single DNN on each benchmark's test set was 0.48cm, 0.75cm, and 10.95cm for the Speaker, Robot, and Adolescent datasets respectively.

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.

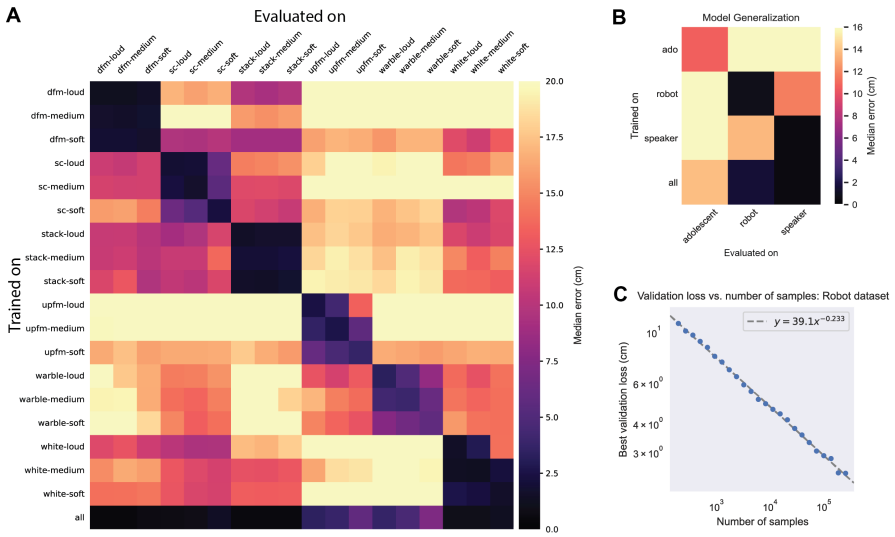

Figure 3: Generalizability across stimulus types and benchmark datasets. A.) Performance of models trained on single stimuli from Speaker dataset and evaluated on all other stimulus types. B.) Performance of models trained on single benchmark datasets and evaluated on all other datasets. C.) Relationship between training set size and test set performance.

To explore this, we performed two experiments. First, to test the ability of deep networks to generalize to new vocal calls with different acoustic features, we partitioned the Speaker Dataset according to stimulus type, trained a deep neural network on each subset, and measured its performance on every stimulus type individually (Fig. 3A). 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. (Fig. 3A, 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 [8, 13].

In a second experiment, we investigated how networks trained on each benchmark dataset to generalize to other datasets. Similar to the results above, we found that performance degraded when models were trained and tested on different datasets (Fig. 3B). Training simultaneously on all datasets improved performance, but performance remained relatively poor on the Adolescent dataset (Fig. 3B, bottom row). The uncertainty calibration also degraded when ensembles were trained and tested on different datasets (See Supplemental Figure 2).

4 Discussion

We have curated and released the first large-scale benchmark datasets for vocal call localization in social rodents. There are several unique challenges associated with this application including the effects of reverberation in small environments and the fine-scale spatial resolution needed to attribute vocal calls to animals congregating within 1-10 cm of each other. Classical approaches appear to work well in acoustically transparent environments [2–4], but we found that deep networks outperformed these approaches in the presence of reverberations. An outstanding challenge is whether these deep learning approaches can be made robust to shifts in the underlying data distribution (see Fig. 3). It is possible that training on larger and more diverse datasets will result in improved outcomes. To explore this, we trained models on subsets of the Robot dataset and plotted the relationship between training set size and test set performance (Fig. 3C). Mirroring findings in the broader literature [14], we found that performance improved according to a power law and showed no clear signs of saturation. Thus, our work suggests that deep networks are a plausible, non-invasive approach to localizing vocal calls in naturalistic rodent interactions, and provides benchmarks and baselines for the community to build upon.

References

[1] Cory T. Miller, David Gire, Kim Hoke, Alexander C. Huk, Darcy Kelley, David A. Leopold, Matthew C. Smear, Frederic Theunissen, Michael Yartsev, and Cristopher M. Niell. "Natural behavior is the language of the brain". In: *Current Biology* 32.10 (2022), R482–R493. ISSN: 0960-9822.

[2] Joshua P Neunuebel, Adam L Taylor, Ben J Arthur, and SE Roian Egnor. "Female mice ultrasonically interact with males during courtship displays". In: *eLife* 4 (May 2015). Ed. by Peggy Mason, e06203. ISSN: 2050-084X.

[3] Max L Sterling, Ruben Teunisse, and Bernhard Englitz. "Rodent ultrasonic vocal interaction resolved with millimeter precision using hybrid beamforming". In: *eLife* 12 (July 2023). Ed. by Brice Bathellier, e86126. ISSN: 2050-084X.

[4] 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.

[5] Pierre-Amaury Grumiaux, Srđan Kitić, Laurent Girin, and Alexandre Guérin. "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.

[6] Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D'Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. "SLEAP: A deep learning system for multi-animal pose tracking". In: *Nature Methods* 19.4 (Apr. 2022), pp. 486–495. ISSN: 1548-7105.

[7] Elsa Steinfath, Adrian Palacios-Muñoz, Julian R Rottschäfer, Deniz Yuezak, and Jan Clemens. "Fast and accurate annotation of acoustic signals with deep neural networks". In: *eLife* 10 (Nov. 2021). Ed. by Ronald L Calabrese, SE Roian Egnor, and Todd Troyer, e68837. ISSN: 2050-084X.

[8] Ralph E Peterson, Aman Choudhri, Catalin Mitelut, Aramis Tanelus, Athena Capo-Battaglia, Alex H Williams, David M Schneider, and Dan H Sanes. "Unsupervised discovery of family specific vocal usage in the Mongolian gerbil". In: *bioRxiv* (2023).

[9] Ella Z Lattenkamp, Sonja C Vernes, and Lutz Wiegrebe. "Volitional control of social vocalisations and vocal usage learning in bats". In: *Journal of Experimental Biology* 221.14 (2018), jeb180729.

[10] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*. 2016, p. 125.

[11] David W. Romero, Erik J. Bekkers, Jakub M. Tomczak, and Mark Hoogendoorn. *Wavelet Networks: Scale Equivariant Learning From Raw Waveforms*. 2020.

[12] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (2017).

[13] Alison J Barker, Grigorii Veviurko, Nigel C Bennett, Daniel W Hart, Lina Mograby, and Gary R Lewin. "Cultural transmission of vocal dialect in the naked mole-rat". In: *Science* 371.6528 (2021), pp. 503–507.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. *Scaling Laws for Neural Language Models*. 2020.

[15] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. "Language Modeling with Gated Convolutional Networks". In: *CoRR* abs/1612.08083 (2016).

Supplementary Information

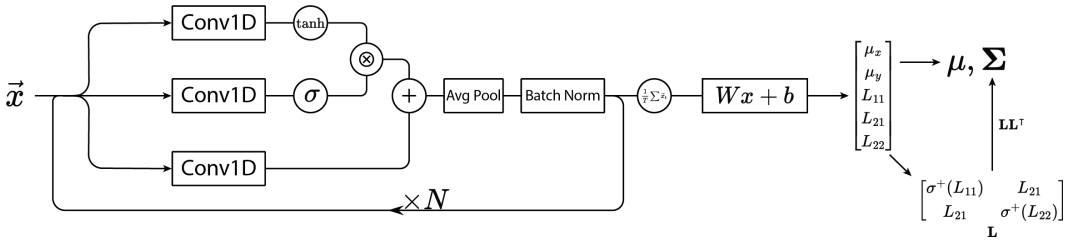

Figure 1: Network architecture.

1 Model Architecture and Approach

The network consists of many convolutional blocks connected in series (Supplementary Figure 1). Each block performs three convolutions of the input data which share hyperparameters (given in Supplementary Table 1) but not weights. Mirroring gated linear units [15] and WaveNet [10], we apply tanh and sigmoid nonlinearities to two of the output tensors and then multiply them with each other element-wise. We add this product to the result of the 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. 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 softmax nonlinearity to the diagonal elements.

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.

2 Training Procedure

During the training phase, we minimize the Negative Log Likelihood of the ground truth labels within the predicted posterior distributions. To achieve this, we employ Stochastic Gradient Descent (SGD) with a momentum of 0.7 and initialize the learning rate at 0.03. 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.

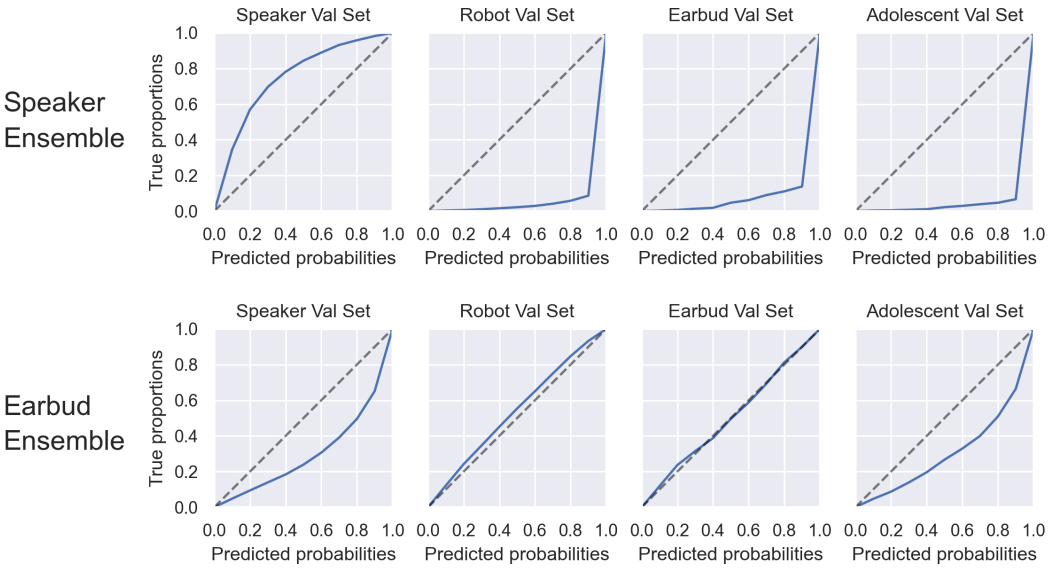

Figure 2: Ensemble calibration differs across datasets.

3 Related work

We compare our model to the Mouse Ultrasonic Source Estimation (MUSE) tool developed to localize ultrasonic vocalizations in anechoic environments [2]. This method works by first computing the cross-correlation between pairs of microphone signals via their frequency-domain representations: $r_{ik}(t) = \int V_i(\omega)V_k(\omega)^* e^{j\omega t} d\omega$. From this, the reduced steered response power (RSRP) can be computed for a hypothesized sound source location by evaluating the following sum over all unique pairings of distinct microphones: $\sum_{i=1}^{M-1} \sum_{k=i+1}^{M} r_{ik}(\Delta\tau_{ik})$ where $\Delta\tau_{ik}$, the time-of-arrival difference of the sound wave between the two microphones, is computed by dividing the difference in the test point's distances to both microphones by the speed of sound. Finally, MUSE searches for the maximal value of the RSRP by evaluating it across a dense grid of test points in the region of interest. This procedure is known as delay-and-sum beamforming.