VAST : The Virtual Acoustic Space Traveler Dataset

Clément Gaultier, Saurabh Kataria, Antoine Deleforge

VAST : The Virtual Acoustic Space Traveler Dataset

Clément Gaultier*, Saurabh Kataria†, and Antoine Deleforge*

*Inria Rennes - Bretagne Atlantique, France (firstname.name@inria.fr)
†Indian Institute of Technology Kanpur, India (saurabhk@iitk.ac.in)

Abstract. This paper introduces a new paradigm for sound source localization referred to as virtual acoustic space traveling (VAST) and presents a first dataset designed for this purpose. Existing sound source localization methods are either based on an approximate physical model (physics-driven) or on a specific-purpose calibration set (data-driven). With VAST, the idea is to learn a mapping from audio features to desired audio properties using a massive dataset of simulated room impulse responses. This virtual dataset is designed to be maximally representative of the potential audio scenes that the considered system may be evolving in, while remaining reasonably compact. We show that virtually-learned mappings on this dataset generalize to real data, overcoming some intrinsic limitations of traditional binaural sound localization methods based on time differences of arrival.

Keywords: Sound Localization, Binaural Hearing, Room Simulation, Machine Learning

1 Introduction

Human listeners have the stunning ability to understand complex auditory scenes using only two ears, *i.e.*, with binaural hearing. Advanced tasks such as sound source direction and distance estimation or speech deciphering in multi-source, noisy and reverberant environments are performed daily by humans, while they are still a challenge for artificial (two-microphone) binaural systems. The main line of research in machine binaural source localization along the past decades has been to estimate the time-difference of arrival (TDOA) of the signal of interest at the two microphones. An estimated TDOA can be approximately mapped to the azimuth angle of a frontal source if the distance between microphones is known, assuming free-field[1] and far-field[2] conditions. Two important limits of these assumptions can be identified. First, they are both violated in most practical scenarios. In the example of an indoor binaural hearing robot, users are typically likely to engage interaction in both far- and near-fields and non-direct sound paths exist due to reflection and diffusion on walls, ceiling, floor, other objects in the room and the robot itself. Second, the intrinsic symmetries of a free-field/far-field binaural system restrict any geometrical estimation to

[1] Free-field means that the sound propagates from the source to the microphones through a single direct path, without interfering objects or reverberations.

[2] Far-field means that the source is placed far enough (*e.g.* > 1.8 meters [12]) from the receiver so that the effect of distance on recorded audio features is negligible.

that of a frontal azimuth angle. Hence, 3D source position (azimuth, elevation, distance) is out of reach in this scope, let alone additional properties such as source orientation, receiver position or room shape.

To overcome intrinsic limitations of TDOA, richer binaural features have been investigated. These include frequency-dependent phase and level differences [20,4], spectral notches [14,7] or the direct-to-reverberant ratio [10]. To overcome the free-field/far-field assumptions, advanced mapping techniques from these features to audio scene properties have been considered. These mapping techniques divide in two categories. The first one is *physics-driven*, *i.e.*, the mapping is inferred from an approximate sound propagation model such as the Woodworth's spherical head formula [20] or the full wave-propagation equation [9]. The second category of mapping is *data-driven*. This approach is sometimes referred to as *supervised sound source localization* [19], or more generally *acoustic space learning* [2]. These methods bypass the use of an explicit, approximate physical model by directly learning a mapping from audio features to audio properties using manually recorded training data [19,4]. They generally yield excellent results, but because obtaining sufficient training data is very time consuming, they only work for a specific room and setup and are hard to generalize in practice. Unlike artificial systems, human listeners benefit from years of adaptive auditory learning in a multitude of acoustic environments. While machine learning recently showed tremendous success in the field of speech recognition using massive amounts of annotated data, equivalent training sets do not exist for audio scene geometry estimation, with only a few specialized manually annotated ones [2,4]. Interestingly, a recent data-driven method [13] used both real and simulated data to estimate room acoustic parameters and improve speech recognition performance, although it was not designed for sound localization.

We propose here a new paradigm that aims at making the best of physics-driven and data-driven approaches, referred to as *virtual acoustic space traveling*. The idea is to use a physics-based room-acoustic simulator to generate arbitrary large datasets of room-impulse responses corresponding to various acoustic environments, adapted to the physical audio system considered. Such impulse responses can be easily convolved with natural sounds to generate a wide variety of audio scenes including *cocktail-party* like scenarios. The obtained corpus can be used to learn a mapping from audio features to various audio scene properties using, *e.g.*, deep learning or other non-linear regression methods [3]. The *virtually-learned* mapping can then be used to efficiently perform real-world auditory scene analysis tasks with the corresponding physical system. Inspired by the idea of an artificial system learning to hear by exploring virtual acoustic environments, we name this proposal the *Virtual Acoustic Space Traveler* (VAST) project. We initiate it by publicly releasing a dedicated project page : `http://theVASTproject.inria.fr` and a first example of VAST dataset. This paper details the guidelines and methodology that were used in the process of building this training set. It then demonstrates that virtually-learned mappings can generalize to real-world test sets, overcoming intrinsic limitations of TDOA-based sound source localization methods.

2 Dataset Design

2.1 General Principles

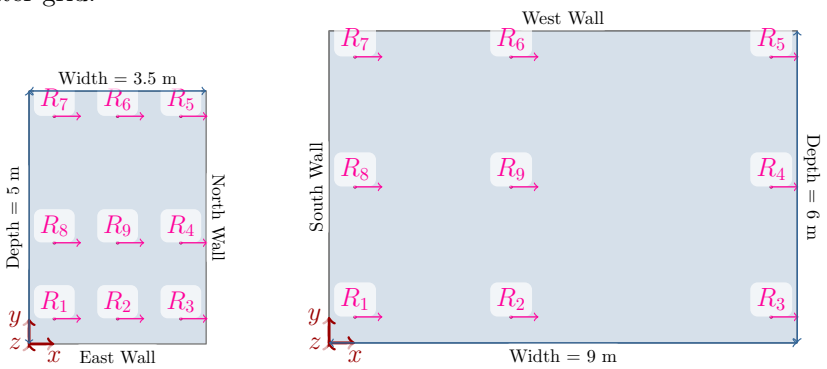
The space of all possible acoustic scenes is vast. Therefore, some trade-offs between the size and the representativity of the dataset must be made when building a training corpus for audio scene geometry estimation. During the process of designing the dataset, we imposed on ourselves the following guidelines:

- The dataset should consist of room impulse responses (RIR). This is a more generic representation than, *e.g.*, specific audio features or audio scenes involving specific sounds. Each RIR should be annotated by all the source, receiver and room properties defining it.
- Virtual acoustic space traveling aims at building a dataset for a **specific audio system** in a variety of environments. Following this idea, some intrinsic properties of the receiver such as its distance to the ground and its head-related transfer functions are kept fixed throughout the simulations. For this first dataset, called *VAST_KEMAR_0*, we chose the emblematic KEMAR acoustic dummy-head, whose measured HRTFs are publicly available. It was placed at 1.70 from the ground, the average human's height.
- We are interested in modeling acoustic environments which are typically encountered in an office building, a university, a hotel or a modern habitation. Acoustics of the type encountered in a cathedral, a massive hangar, a recording studio or outdoor are deliberately left aside here. Surface materials and diffusion profiles are chosen accordingly.
- To make the dataset easily manipulable on a simple laptop, we aimed at keeping its total size under 10 GigaBytes. To handle datasets of larger order of magnitudes would require users to have access to specific hardware and software which is not desired here. *VAST_KEMAR_0* measures 6.4 GB.

2.2 Room Simulation and Data Generation

The efficient C++/MATLAB "shoebox" 3D acoustic room simulator ROOMSIM developed by Schimmel et al. is selected for simulations [16]. This software takes as input a room dimension (width, depth and height), a source and receiver position and orientation, a receiver's head-related-transfer function (HRTF) model, and frequency-dependent absorption and diffusion coefficients for each surface. It outputs a corresponding pair of room impulse responses (RIR) at each ear of the binaural receiver. Specular reflections are modeled using the image-source method [1], while diffusion is modeled using the so-called *rain-diffusion* algorithm. In the latter, sound rays uniformly sampled on the sphere are sent from the emitter and bounced on the walls according to specular laws, taking into account surface absorption. At each impact, each ray is also randomly bounced towards the receiver with a specified probability (the frequency-dependent *diffusion coefficient* of the surface). The total received energy at each frequency is then aggregated using histograms. This model was notably showed to realistically account for sound scattering due to the presence of objects, by comparing simulated RIRs with measured ones in [22]. The study [8] suggests that such diffusion

effects play an important role in sound source localization. *VAST_KEMAR_0* contains over 110,000 RIR, which required about 700 CPU-hours of computation. This was done using a massively parallelized implementation on a large computer grid.



(a) Small Room (Height = 2.5m) (b) Large Room (Height = 3.5m)

Fig. 1. Top views of training rooms with receiver positions and orientations.

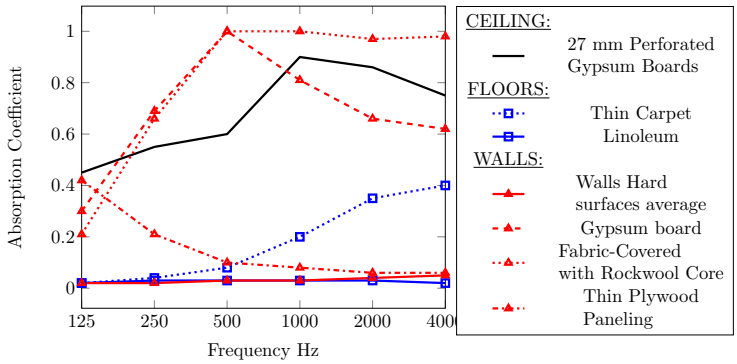


Fig. 2. Absorption Profiles

2.3 Room Properties: Size and Surfaces

An obvious choice to generate virtual rooms with maximal variability would be to draw a random room size and random frequency-dependent absorption and diffusion profiles of surfaces for each generated RIR. This approach however, has several drawbacks. First, it makes impossible the generation of realistic audio scenes containing several sources, for which the receiver position and the room must be fixed. Second, the space of possible rooms is so vast that reliably sampling it at random is unrealistic. Third, changing source, receiver and room parameters all at the same time prevents from getting insights on the individual influence of these parameters. On the other hand, sampling all combinations of parameters in an exhaustive way quickly leads to enormous data size. As a trade-off, we designed 16 realistic rooms representative of typical reverberation time (RT_{60}) and surface absorption profiles encountered in modern buildings. Two room sizes were considered: a small one corresponding to a typical office

Table 1. Description of simulated training rooms in VAST

Room Number	Floor	Ceiling	Walls	Width [m]	Depth [m]	Height [m]
1	Thin Carpet	Perforated 27 mm gypsum board	Walls Hard Surfaces Average	9	6	3.5
2	Thin Carpet	Perforated 27 mm gypsum board	Gypsum Board with Mineral Filling	9	6	3.5
3	Thin Carpet	Perforated 27 mm gypsum board	Fabric-Covered Panel with Rockwool Core	9	6	3.5
4	Thin Carpet	Perforated 27 mm gypsum board	Thin Plywood Paneling	9	6	3.5
5	Linoleum	Perforated 27 mm gypsum board	Walls Hard Surfaces Average	9	6	3.5
6	Linoleum	Perforated 27 mm gypsum board	Gypsum Board with Mineral Filling	9	6	3.5
7	Linoleum	Perforated 27 mm gypsum board	Fabric-Covered Panel with Rockwool Core	9	6	3.5
8	Linoleum	Perforated 27 mm gypsum board	Thin Plywood Paneling	9	6	3.5
9	Thin Carpet	Perforated 27 mm gypsum board	Walls Hard Surfaces Average	3.5	5	2.5
10	Thin Carpet	Perforated 27 mm gypsum board	Gypsum Board with Mineral Filling	3.5	5	2.5
11	Thin Carpet	Perforated 27 mm gypsum board	Fabric-Covered Panel with Rockwool Core	3.5	5	2.5
12	Thin Carpet	Perforated 27 mm gypsum board	Thin Plywood Paneling	3.5	5	2.5
13	Linoleum	Perforated 27 mm gypsum board	Walls Hard Surfaces Average	3.5	5	2.5
14	Linoleum	Perforated 27 mm gypsum board	Gypsum Board with Mineral Filling	3.5	5	2.5
15	Linoleum	Perforated 27 mm gypsum board	Fabric-Covered Panel with Rockwool Core	3.5	5	2.5
16	Linoleum	Perforated 27 mm gypsum board	Thin Plywood Paneling	3.5	5	2.5
0	Anechoic room					

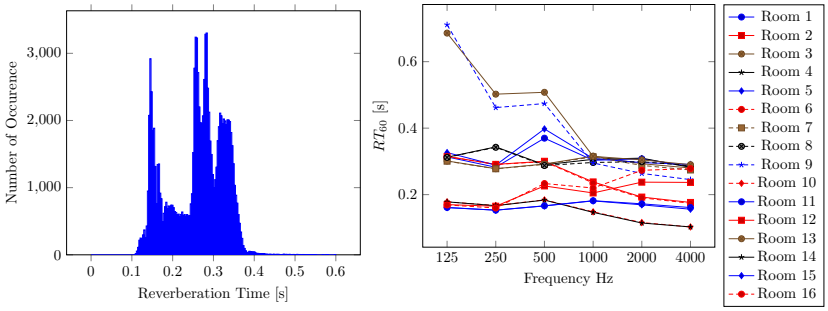
or bed room (Fig. 1(a)), and a larger one corresponding to a lecture or entrance hall (Fig. 1(b)). For each room, floor, ceiling and wall materials which are representative in terms of absorption profile and are commonly encountered in nowadays buildings were chosen from [21]. The graph on Fig. 2 displays the absorption profiles of the selected materials, namely, 4 for the walls, 2 for the floor and 1 for the ceiling. The gypsum board material chosen for the ceiling was kept fixed throughout the dataset, as it represents well typical ceiling absorption profiles [21]. "Walls hard surface average" is in fact an average profile over many surfaces such as brick or plaster [21]. Combining all possible floors, walls and room sizes yielded the 16 rooms listed in Table 1.

Importantly, typical rooms also contain furniture and other objects responsible for random sound scattering effects, *i.e.*, diffusion. Following the acoustic study in [5], a unique frequency-dependent diffusion profile was used for all surfaces. The chosen profile is the average of the 8 configurations measured in [5], corresponding to varying numbers of chairs, table, computers and people in a room. Both absorption and diffusion profiles are piecewise-linearly interpolated from 8 Octave bands from 125 Hz to 4 kHz.

2.4 Reverberation Time

A common acoustic descriptor for rooms is the reverberation time (RT_{60}). Figure 3(a) displays the estimated RT_{60} distribution across the VAST Training Dataset. Fig. 3(b) shows the RT_{60} for each room by octave band. RT_{60}'s were estimated from the room impulse responses following the recommendations in [17]. From these estimations, we decided to crop the room impulse responses provided in the datasets above the RT_{60}, with a 30 ms margin. This technique allows to shrink the dataset while keeping data points of interest and discarding the rest. To further complies with memory limitations, we chose to encode the room impulse response samples with single floats (16 bit). As can be seen in Fig. 3 the 16 chosen rooms present a quite good variability in terms of reverberation times in the range 100ms-400ms. Larger RT_{60} of the order of 1 second could be obtain

by using highly reflective materials on all surfaces, creating an echo chamber. However, this rarely occurs in realistic buildings.



(a) Global RT_{60} distribution (b) Octave bands RT_{60}

Fig. 3. Reverberation Time.

2.5 Source and Receiver Positions

A relatively poorly-studied though important effect in sound source localization is the influence of the receiver's position in the room, especially its distance to the nearest surface. In order to accurately capture this effect, 9 receiver positions are used for each of the 16 rooms, while the height of the receiver is fixed at 1.7 m. Figure 1 shows top views of the rooms with receiver positions. Positions from R_1 to R_8 are set 50 cm from the nearest wall(s) whereas R_9 is approximately placed in the middle of the room. Perfectly symmetrical configurations are avoided to make the dataset as generic as possible, without singularities. The receiver is always facing the north wall as a convention. For each of the 9 receiver positions, sources are placed on spherical grids centered on the receiver. Each sphere consists of regularly-spaced elevation lines each containing sources at regularly-spaced azimuths, with a spacing of 9°. The equator elevation line and the first azimuth angle of each line are randomly offset by $-4.5°$ to $+4.5°$ in order to obtain a dense sphere sampling throughout the dataset. Six spherical grid radii are considered, yielding source distances of 1, 1.5, 2, 3, 4 and 6 meters. Sources falling outside of the room or less than 20cm from a surface are removed.

2.6 Test Sets

To test the generalizability of mappings learned on the *VAST_KEMAR_0* dataset, we built four simulated test sets differing from the training dataset on various levels. A first challenge is to test robustness to random positioning, since the training set is built with regular spherical source grids and fixed listener positions. Hence, the 4 testing sets contain completely random source and receiver positions in the room. Only the receiver's height is fixed to 1.7m, and both receiver and source are set within a 20 cm safety margin within the room boundaries. Test sets 2 and 4 feature random receiver orientation (yaw angle), as opposed to the receiver facing north in the training set. Test 1 and 2 contain 1,000 binaural RIRs (BRIRs) for each of the 16 rooms of Table 1. Finally, test sets 3 and 4 contain 10,000 BRIRs, each corresponding to a random room size

(walls from $3m \times 2m$ to $10m \times 4m$) and random absorption properties of walls and floor picked from Fig. 2. Different surfaces for all 4 walls are allowed.

In addition to these simulated test sets, three binaural RIR datasets recorded with the KEMAR dummy head in real rooms have been selected, as listed below:

- **Auditorium 3 [11]** was recorded at TU Berlin in 2014 in a trapezium-shaped lecture room of dimensions 9.3m × 9m and $RT_{60} \approx 0.7$s. 3 individual sources placed 1.5m from the receiver at different azimuth and 0° elevation were recorded. For each source, one pair of binaural RIR is recorded for each receivers' head yaw angle from −90° to +90°, with 1° steps.
- **Spirit [11]** was recorded at TU Berlin in 2014 in a small rectangular office room of size 4.3m × 5m, $RT_{60} \approx 0.5$s, containing various objects, surfaces and furniture near the receiver. The protocol is the same as Auditorium 3 except sources are placed 2m from the receiver.
- **Classroom [18]** was recorded at Boston University in 2005 in a 5m × 9m × 3.5m carpeted classroom with 3 concrete walls and one sound-absorptive wall ($RT_{60} = 565$ms). The receiver is placed in 4 locations of the room including 3 with at least one nearby wall.

Note that the KEMAR HRTF measurements used to simulate the VAST dataset was recorded by yet another team, in MIT's anechoic chamber in 1994, as described in [6].

3 Virtually Supervised Sound Source Localization

For all experiments in this section, all training and test sets used are reduced to contain only frontal sources (azimuth in $[-90°, +90°]$) with elevation in [−45°, +45°] and distances between 1 and 3 meters. As mentioned in the introduction, sound source localization consists in two steps: calculating auditory features from binaural signals followed by mapping these features to a source position. Robustly estimating features can be difficult when dealing with additive noise, sources with sparse spectra such as speech or music, and source mixtures. We leave this problematic aside in this paper, and focus on mapping clean features to source positions. Hence, we use *ideal* features directly calculated from the clean room impulse responses in all experiments.

We first make an experiment to put forward some intrinsic limitations of TDOA-based azimuth estimation. Fig. 4 plots TDOAs against the source's azimuth angle for different subsets of VAST. TDOAs (in samples) were computed as the integer delay in $[-15, +15]$ maximizing the correlation between the first 500 samples of the left and the right impulse responses. As can be seen in Fig. 4(a), a near-linear relationship between frontal azimuth and TDOA exists in the anechoic case, regardless of the elevation. This matches previously observed results in binaural sound localization [20,15,4]. When the receiver is placed in the middle of the 16 reverberant rooms, (Fig. 4(b)), some outliers appear due to reflections. This effect is dramatically increased when the receiver is placed 50 centimeters from a wall (Fig. 4(c) and 4(d)), where stronger early reflections

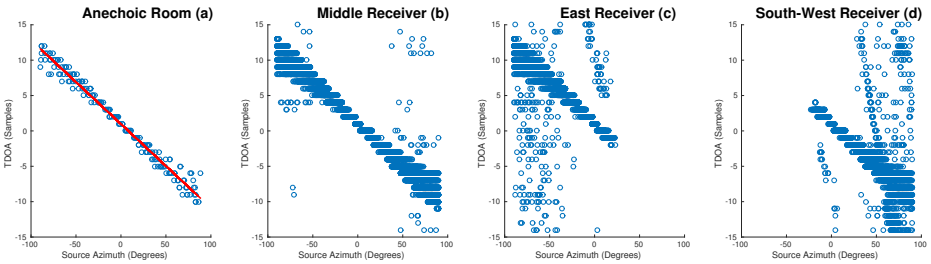


Fig. 4. TDOA as a function of source azimuth in various settings

Table 2. Azimuth absolute estimation errors in degrees with 3 different methods, showed in the form $avg \pm std(out\%)$, where *avg* and *std* denote the mean and standard deviation of *inlying* absolute errors ($< 30°$) while out denotes the percentage of outliers.

Test data ↓	TDOA	GLLiM (Anech. train.)	GLLiM (VAST train.)
VAST Testing Set 1	$5.49 \pm 4.6(5.6\%)$	$8.63 \pm 7.6(12\%)$	$4.38 \pm 4.9(1.8\%)$
VAST Testing Set 2	$5.37 \pm 4.4(6.0\%)$	$8.09 \pm 7.5(12\%)$	$4.32 \pm 4.7(1.6\%)$
VAST Testing Set 3	$5.21 \pm 4.5(4.6\%)$	$8.46 \pm 7.5(5.2\%)$	$4.23 \pm 4.4(1.8\%)$
VAST Testing Set 4	$5.14 \pm 4.4(3.3\%)$	$8.21 \pm 7.2(4.8\%)$	$4.25 \pm 4.4(0.6\%)$
Auditorium 3 [11]	$7.02 \pm 4.7(1.4\%)$	$8.01 \pm 7.0(5.9\%)$	$5.03 \pm 4.5(0.0\%)$
Spirit [11]	$5.19 \pm 3.4(0.0\%)$	$12.2 \pm 8.3(15\%)$	$4.50 \pm 5.6(0.4\%)$
Classroom [18]	$5.71 \pm 3.7(3.7\%)$	$9.47 \pm 7.3(5.2\%)$	$6.50 \pm 5.9(0.0\%)$

are present. This suggests that the TDOA, even when ideally estimated, is not adapted to binaural sound source localization in realistic indoor environments.

We then compare azimuth estimation errors obtained with the TDOA-based method described above, a learning-based method trained on anechoic HRTF measurements (Room 0), and a learning-based method trained on VAST, using the 4 simulated and 3 real test sets described in Section 2.6. TDOAs were mapped to azimuth values using the affine regression coefficients corresponding to the red line in Fig. 4(a). The chosen learning-based sound source localization method is the one described in [4]. It uses Gaussian Locally Linear Regression (GLLiM, [3]) to map high-dimensional feature vectors containing frequency-dependent interaural level and phase differences from 0 to 8000 Hz to low-dimensional source positions. In our case, the GLLiM model with K locally-linear components was trained on N interaural feature vectors of dimension $D = 1537$ associated to 3-dimensional source positions in spherical coordinate (azimuth, elevation and

Table 3. Elevation and distance absolute estimation errors obtained with GLLiM trained on VAST. Outliers correspond to errors larger than 15° or 1m.

Test data ↓	Elevation (°)	Distance (m)
VAST Testing Set 1	$5.91 \pm 4.1(23\%)$	$0.43 \pm 0.3(19\%)$
VAST Testing Set 2	$6.05 \pm 4.2(27\%)$	$0.44 \pm 0.3(20\%)$
VAST Testing Set 3	$6.05 \pm 4.1(27\%)$	$0.43 \pm 0.3(21\%)$
VAST Testing Set 4	$6.03 \pm 4.2(26\%)$	$0.44 \pm 0.3(21\%)$
Auditorium 3 [11]	$7.92 \pm 4.4(44\%)$	$0.45 \pm 0.3(23\%)$
Spirit [11]	$7.44 \pm 4.3(30\%)$	$0.52 \pm 0.3(25\%)$
Classroom [18]	$8.40 \pm 4.1(45\%)$	$0.41 \pm 0.3(6.5\%)$

distance). $K = 8$ components were used for the anechoic training set ($N = 181$) and $K = 100$ for the (reduced) VAST dataset ($N \approx 41,000$). All 3 methods showed comparably low testing computational times, in the order of 10ms for 1 second of input signal. Table 2 summarizes obtained azimuth estimation errors. As can be seen, the learning method trained on VAST outperforms the two others on all datasets, with significantly less outliers and a globally reduced average error of inliers. This is encouraging considering the variety of testing data used. In addition, Table 3 shows that GLLiM trained on VAST is capable of approximately estimating the elevation and distance of the source, which is known to be particularly difficult from binaural data. While elevation estimation on real data remains a challenge, results obtained on simulated sets are promising.

4 Conclusion

We introduced the new concept of virtual acoustic space traveling and released a first dataset dedicated to it. A methodology to efficiently design such a dataset was provided, making extensions and improvements of the current version easily implementable in the future. Results show that a learning-based sound source localization method trained on this dataset yields better localization results than when trained on anechoic HRTF measurements, and performs better than a TDOA-based approach in azimuth estimation while being able to estimate source elevation and distance. To the best of the authors' knowledge, this is the first time a sound localization method trained on simulated data is successfully used on real data, validating the new concept of virtual acoustic space traveling. The learning approach could still be significantly improved by considering other auditory features, by better adapting the mapping technique to spherical coordinates and by annotating training data with further acoustic information. Other learning methods such as deep neural networks may also be investigated.

References

1. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America 65(4), 943–950 (1979)
2. Deleforge, A., Forbes, F., Horaud, R.: Acoustic space learning for sound-source separation and localization on binaural manifolds. International journal of neural systems 25(01), 1440003 (2015)
3. Deleforge, A., Forbes, F., Horaud, R.: High-dimensional regression with gaussian mixtures and partially-latent response variables. Statistics and Computing 25(5), 893–911 (2015)
4. Deleforge, A., Horaud, R., Schechner, Y.Y., Girin, L.: Co-localization of audio sources in images using binaural features and locally-linear regression. IEEE Transactions on Audio, Speech, and Language Processing 23(4), 718–731 (2015)
5. Faiz, A., Ducourneau, J., Khanfir, A., Chatillon, J.: Measurement of sound diffusion coefficients of scattering furnishing volumes present in workplaces. In: Acoustics 2012 (2012)
6. Gardner, W.G., Martin, K.D.: Hrtf measurements of a kemar. The Journal of the Acoustical Society of America 97(6), 3907–3908 (1995)

7. Hornstein, J., Lopes, M., Santos-Victor, J., Lacerda, F.: Sound localization for humanoid robots-building audio-motor maps based on the hrtf. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1170–1176. IEEE (2006)
8. Kataria, S., Gaultier, C., Deleforge, A.: Hearing in a shoe-box: Binaural source position and wall absorption estimation using virtually supervised learning. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2017)
9. Kitić, S., Bertin, N., Gribonval, R.: Hearing behind walls: localizing sources in the room next door with cosparsity. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3087–3091. IEEE (2014)
10. Lu, Y.C., Cooke, M.: Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. IEEE Transactions on Audio, Speech, and Language Processing 18(7), 1793–1805 (2010)
11. Ma, N., May, T., Wierstorf, H., Brown, G.J.: A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2699–2703. IEEE (2015)
12. Otani, M., Hirahara, T., Ise, S.: Numerical study on source-distance dependency of head-related transfer functions. The Journal of the Acoustical Society of America 125(5), 3253–3261 (2009)
13. Parada, P.P., Sharma, D., Lainez, J., Barreda, D., van Waterschoot, T., Naylor, P.A.: A single-channel non-intrusive c50 estimator correlated with speech recognition performance. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(4), 719–732 (2016)
14. Raykar, V.C., Duraiswami, R., Yegnanarayana, B.: Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. The Journal of the Acoustical Society of America 118(1), 364–374 (2005)
15. Sanchez-Riera, J., Alameda-Pineda, X., Wienke, J., Deleforge, A., Arias, S., Čech, J., Wrede, S., Horaud, R.: Online multimodal speaker detection for humanoid robots. In: 2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012). pp. 126–133. IEEE (2012)
16. Schimmel, S.M., Muller, M.F., Dillier, N.: A fast and accurate "shoebox" room acoustics simulator. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 241–244. IEEE (2009)
17. Schroeder, M.R.: New method of measuring reverberation time. The Journal of the Acoustical Society of America 37(3), 409–412 (1965)
18. Shinn-Cunningham, B.G., Kopco, N., Martin, T.J.: Localizing nearby sound sources in a classroom: Binaural room impulse responses. The Journal of the Acoustical Society of America 117(5), 3100–3115 (2005)
19. Talmon, R., Cohen, I., Gannot, S.: Supervised source localization using diffusion kernels. In: 2011 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA). pp. 245–248. IEEE (2011)
20. Viste, H., Evangelista, G.: On the use of spatial cues to improve binaural source separation. In: Proceedings of 6th International Conference on Digital Audio Effects (DAFx-03). pp. 209–213. No. LCAV-CONF-2003-026 (2003)
21. Vorländer, M.: Auralization: fundamentals of acoustics, modeling, simulation, algorithms and acoustic virtual reality. Springer Science & Business Media (2007)
22. Wabnitz, A., Epain, N., Jin, C., Van Schaik, A.: Room acoustics simulation for multichannel microphone arrays. In: Proceedings of the International Symposium on Room Acoustics. pp. 1–6 (2010)