Hearing in a shoe-box : binaural source position and wall absorption estimation using virtually supervised learning

Saurabh Kataria, Clément Gaultier, Antoine Deleforge

HEARING IN A SHOE-BOX : BINAURAL SOURCE POSITION AND WALL ABSORPTION ESTIMATION USING VIRTUALLY SUPERVISED LEARNING

Saurabh Kataria[⋆†] *Clément Gaultier*[⋆] *Antoine Deleforge*[⋆]

[⋆] Inria Rennes - Bretagne Atlantique, France [†] Indian Institute of Technology Kanpur, India

ABSTRACT

This paper introduces a new framework for supervised sound source localization referred to as virtually-supervised learning. An acoustic shoe-box room simulator is used to generate a large number of binaural single-source audio scenes. These scenes are used to build a dataset of spatial binaural features annotated with acoustic properties such as the 3D source position and the walls' absorption coefficients. A probabilistic high- to low-dimensional regression framework is used to learn a mapping from these features to the acoustic properties. Results indicate that this mapping successfully estimates the azimuth and elevation of new sources, but also their range and even the walls' absorption coefficients solely based on binaural signals. Results also reveal that incorporating random-diffusion effects in the data significantly improves the estimation of all parameters.

Index Terms— Sound source localization, Acoustic Modeling, Machine Learning

1. INTRODUCTION

Most existing methods in multichannel audio signal processing, including speech enhancement, denoising or source separation, rely on a good knowledge of the *geometry* of the audio scene. In other words, what are the positions of the sources, sensors, and how does the sound propagate between them. Since this knowledge is most of the time unavailable, it is usually estimated from measured signals. A typical assumption is the free-field model in which the sound propagates in a straight line from each source to each sensor. Then, if sensor positions are known, sound source directions may be determined based on estimated time-differences of arrival. However, typical real-world audio scenes include reflecting and diffusive walls, floor, ceiling or filtering effects due, *e.g.*, to the head of a binaural (2 microphones) receiver. Accurate audio scene geometry estimation is much more challenging in this context.

Two orthogonal research directions have recently emerged to tackle this challenge. The first one is *physics-driven*, and consists in using more advanced acoustical models of the audio scenes. Such models may range from the image source model that incorporates specular reflections [1], to the full wave propagation equation within boundaries of arbitrary shape and impedance [2, 3]. These methods are computationally intensive but yield encouraging results in simulated settings. The second direction is *data-driven*, and consists in learning a mapping from measured high-dimensional acoustic features to source postions. Such mappings are learned from carefully recorded datasets in a supervised [4, 5] or semi-supervised [6] way. Since obtaining these datasets is time consuming, the methods are usually working well for one specific room and setup, and are hard to generalize in practice.

We now propose a third direction that somehow makes use of both worlds, namely, *virtually supervised learning*. The idea is to use a physics-based room-acoustic simulator to generate arbitrary large datasets of audio-features in various geometrical settings. These data are then used to learn an efficient mapping from audio features to geometrical parameters. In this study, we make a first proof-of-concept by focusing on the scenario of a binaural receiver in a shoebox room of specific size. Over $80,000$ audio scenes with varying source direction, source distance, wall absorption, and random diffusion are generated using the ROOMSIM software of Shimmel and et al. [7]. We then extend the supervised sound-source localization method of [5] to not only estimate 2D directions (azimuth, elevation) but also source ranges and mean wall absorption coefficients, solely based on binaural signals. Our experiments show promising results, and reveal that in the considered setting, the addition of diffusion effects significantly improve estimation of all parameters. While diffusion effects are most often neglected in the sound source localization literature, this suggests that they carry rich spatial information which may be helpful for binaural hearing.

2. DESCRIPTION OF EXPERIMENTAL SETUP

The problem of single-source localization in a reverberant room using a binaural receiver is considered. It is well-known from both psychophysical [8, 9] and machine hearing [5] studies that perceived binaural features do not only depend on the source's azimuth, but also on its elevation, its range, the position of the receiver in the room, and the room acoustic properties. The aim of this study is to investigate whether some of these additional parameters can be learned and estimated based on perceived spatial binaural features.

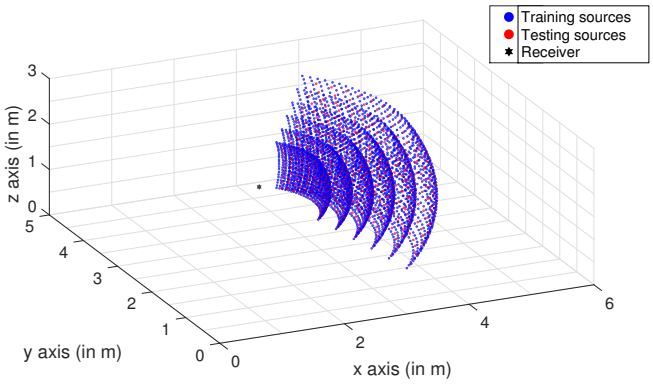


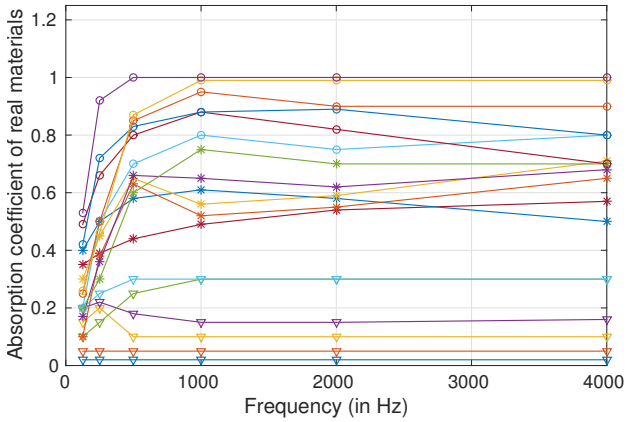
Fig. 1. Illustration of room setup

Room dimensions (in m)	6 x 5 x 3.3
Receiver coordinates (in m)	(2, 2.5, 1.6)
Receiver HRTF model	MIT Kemar [10]
Sensor type	Omnidirectional
Six grid ranges (in m)	{1, 1.3, 1.6, 1.9, 2.2, 2.5}
Angular range for sources (azimuth)	[-45°, 45°]
Angular range for sources (elevation)	[-30°, 30°]
Frequency bins from which absorption and diffusion profiles are linearly interpolated (in KHz)	{0.125, 0.25, 0.5, 1, 2, 4}
Frequency-dependent absorption values for ceiling (gypsum board)	(0.45, 0.55, 0.60, 0.90, 0.86, 0.75)
Frequency-dependent absorption values for floor (thin carpet)	(0.02, 0.04, 0.08, 0.20, 0.35, 0.40)
Frequency-dependent diffusion values for all surfaces	(0.003, 0.004, 0.045, 0.077, 0.210, 0.431)

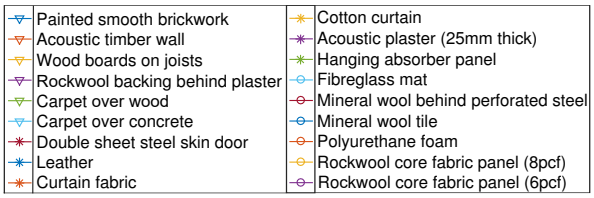
Table 1. General information about room setup.

The number and range of such parameters can be extremely vast considering the variety of real-world rooms, from anechoic chambers to cathedrals. Therefore, we choose to make a trade-off between realism and the number of parameters considered. The room size, receiver position and absorption profiles of the floor and ceiling are assumed fixed in all experiments. However, the azimuth, elevation and range of the source as well as the absorption profile of the walls (assumed identical for all walls), are varied. In addition, random-diffusion effects are added to account for the presence of sound scattering due to objects in the room. Finally, a white-noise emitter is used to avoid biases due to the specific spectral shapes of, *e.g.*, speech signals. Details of the room simulation parameters are showed in Table 1).

The efficient C++/MATLAB "shoebox" 3D acoustic room simulator ROOMSIM developed by Schimmel et al. is selected for simulations[7]. This software takes as input a room dimension (width, depth and height), a source and receiver position, 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 [11], 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



(a) Absorption value of 18 materials with frequency



(b) Legend for the plot above

Fig. 2. Pictorial view of the dataset constructed

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 [12].

The source positions considered are distributed on six spherical grids centered on the receiver, with radii vary from 1 meter to 2.5 meters. Each grid consists of 651 training positions and 150 distinct testing positions with uniformly distributed azimuths and elevations. The angular separation between consecutive positions is 3° for training and 6° for testing. Fig. 1 shows the receiver and the different source positions within the room.

Since the absorption profile of a surface depends on frequency, the number of parameters required to fully model an absorption profile is too high, making its estimation unrealistic. A database of 18 measured absorption profiles of real materials was collected from [13] and [14], and is listed in Fig. 2b. As illustrated in Fig. 2a, these materials show a strong absorption variability below 500 Hz. Then, the absorption slowly decreases or increases with frequency to reach a nearly constant value. These profiles were selected because the standard deviation (std) of their absorption is always below 0.07 between 500 Hz and 4 kHz. This allows to summarize the absorption profile of the four walls with a single parameter: the mean absorption coefficient above 500 Hz. Con-

sequently, observations in frequencies under 500 Hz will be ignored in all experiments. Note that there is a second reason for ignoring these frequencies. The 60 dB reverberation times (RT60) measured in our experiments ranged from 0.09 to 1 second. This gives an upper-bound for the *Schroeder frequency* of 200 Hz. The Schroeder frequency is given by $f_{\text{sch}} = 2000\sqrt{\text{RT60}/V}$ where V is the room volume in m^3, and provides the frequency limit above which sound intensity is approximately homogeneous and isotropic [15, 16]. Above that limit, the modal theory does not hold and is replaced by statistical models of diffuse fields. The simulator, which relies on such models, is thus more likely to give realistic results above that bound. The diffusion profile used in all experiments corresponds to the diffusion field added by three chairs, one table and one computer, as measured in [17]. As in, *e.g.*, [12], the same profile is used for all surfaces. Overall, each pair of generated RIR depends on four parameters: the source's azimuth, elevation and range, and the mean absorption coefficient of walls.

3. MAPPING BINAURAL FEATURES TO GEOMETRICAL PROPERTIES

3.1. Computing Binaural features

Let $\boldsymbol{u} \in \mathbb{R}^4$ be a parameter vector containing the source's azimuth, elevation and range, and the mean wall absorption. We denote the associated generated left and right RIR by $(\boldsymbol{h}^{\text{L}}(\boldsymbol{u}), \boldsymbol{h}^{\text{R}}(\boldsymbol{u}))$. Each of these pairs is convolved with a 1 second random white Gaussian noise signal, and the result is resampled at 8kHz. The short-time Fourier transform is then applied to both signals, using a 64ms sliding time window with 50% overlap. This results in a left-microphone spectrogram $\{L(f,t)\}_{f=1,t=1}^{F,T}$ and a right-microphone spectrogram $\{L(f,t)\}_{f=1,t=1}^{F,T}$, where $F = 256$ and $T = 32$. If $\{S(f,t)\}_{f=1,t=1}^{F,T}$ denotes the emitted white-noise spectrogram, under the assumption that most of the RIR energy is concentrated on the first 64ms, we have the following approximate multiplicative model

$$\begin{cases} L(f,t) \approx \hat{\boldsymbol{h}}^{L}(f,\boldsymbol{u})S(f,t) \\ R(f,t) \approx \hat{\boldsymbol{h}}^{R}(f,\boldsymbol{u})S(f,t) \end{cases} \quad (1)$$

where $\hat{\cdot}$ denotes the discrete Fourier transform. The *interaural level difference* (ILD) and *interaural phase difference* (IPD) spectrograms are defined by

$$\begin{cases} \text{ILD}(f,t) = 20 * \log(|L(f,t)|/|R(f,t)|) \in \mathbb{R} \\ \text{IPD}(f,t) = \frac{L(f,t)/|L(f,t)|}{R(f,t)/|R(f,t)|} \in \mathbb{C} \equiv \mathbb{R}^2. \end{cases} \quad (2)$$

Using the approximation (1), it is easily seen that both ILD and IPD solely depend on the parameter vector $\boldsymbol{u}$ and do not depend on the emitted signal. Similarly to [5], the ILD and IPD spectrograms are vertically concatenated and averaged over time to form a high-dimensional feature vector $\boldsymbol{y} \in \mathbb{R}^D$ associated to the low-dimensional parameter vector $\boldsymbol{u} \in \mathbb{R}^L (L = 4)$. As explained in Section 2, only the $F' = 481$ bins corresponding to frequencies above 500 Hz are used in $\boldsymbol{y}$, resulting in a dimension $D = 3F' = 1443$ in practice.

3.2. Gaussian Locally-Linear Mapping

The training dataset is composed of N pairs $\{(\boldsymbol{y}_n, \boldsymbol{u}_n)\}_{n=1}^N \subset \mathbb{R}^D \times \mathbb{R}^L$. A mapping needs to be learned from this dataset such that given a new test observation $\widetilde{\boldsymbol{y}}_t \in \mathbb{R}^D$, an associated parameter vector $\widetilde{\boldsymbol{u}}_t$ can be estimated. To achieve this, we use the high- to low-dimensional regression method *Gaussian locally-linear mapping* (GLLiM) proposed in [18]. GLLiM is a probabilistic method that estimates K local affine transformation from the space of $\boldsymbol{u}$ to the space of $\boldsymbol{y}$ using a Gaussian mixture model. This mapping is then reversed through Bayes' inversion, yielding an efficient estimator of $\boldsymbol{u}$ given $\boldsymbol{y}$. GLLiM was successfully applied to supervised 2D sound source localization on a real dataset in [5]. In practice, a fixed value $K = 25$ is used in all experiments, as this showed to be a good trade-off between accuracy and computational time using preliminary validation sets.

4. EXPERIMENTATION AND RESULTS

We first conduct an experiment to reproduce the supervised binaural 2D localization results of [5] in our setting. The GLLiM model is trained on 651 individual white noise (WN) recordings obtained from sources lying on a grid at a range of 1 m (the closest grid in Fig. 1). The chosen wall absorption profile is "Rockwool backing behind plaster" (Fig. 2b) with mean absorption value 0.16. Random diffusion effects are added on both training and testing data. Two test cases are compared. First, testing in the same configuration, *i.e.* , absorption and range are the same between training and testing. Second, testing with the different wall absorption profile "Rockwool core fabric panel (8pcf)" (Fig. 2b, mean absorption value 0.96) and a grid range of 2.5 m. As expected, table 2 illustrates that the localization error is higher when testing in a configuration different from the training configuration, in particular in elevation. The same phenomenon was observed in [5] with real data. As in [5], it can be noted that estimating elevation is harder than azimuth, which is expected because of head symmetry. This difficulty seems further increased here by the use of a cutoff frequency at 500 Hz. This fundamental experiment motivates the idea of a robust training using multiple configurations of absorption values and grid ranges.

	Testing in the same configuration	Testing in a different configuration
Azimuthal error (°)	1.67 ± 1.22	1.99 ± 1.42
Elevation error (°)	8.78 ± 7.08	15.79 ± 12.39

Table 2. Comparing the mean and std of absolute localization errors when training with a single room and source range.

Training set annotation →	Direction+Range+Absorption	Direction+Range+Absorption (no diffusion)	Direction Only	Direction+Absorption	Direction+Range
Azimuth (°)	1.78 ± 1.34	2.16 ± 1.62	1.72 ± 1.43	2.00 ± 1.51	1.91 ± 1.52
Elevation (°)	7.87 ± 6.45	11.3 ± 7.95	8.81 ± 7.81	8.45 ± 6.86	9.44 ± 7.55
Range (cm)	54.2 ± 29.65	56.8 ± 34.3	-	-	58.5 ± 32.4
Absorption	0.18 ± 0.14	0.80 ± 0.44	-	0.22 ± 0.17	-

Table 3. Mean±std localization errors for various training sets. The second column is the same as first one but without diffusion. Vertical labels denote variables to be estimated. Horizontal labels denote parameters supplied during training (vector $\boldsymbol{u}$).

We then trained the GLLiM model using all 651 training directions, 6 ranges (Fig. 1), and 21 constant wall absorption values in $\{0, 0.05, ..., 1\}$. This results in a dataset of $N = 651 \times 6 \times 21 = 82,026$ binaural feature vectors associated to $L = 4$-dimensional parameter vectors $\{(\boldsymbol{y}_n, \boldsymbol{u}_n)\}_{n=1}^{N} \subset \mathbb{R}^D \times \mathbb{R}^L$. The 21 absorption values used here correspond to ideal materials with perfectly constant absorption coefficient in the frequency range $[0.5, 4]$ KHz. Testing is done on the 108 test directions (Fig. 1), 6 grid ranges, and the 18 real material absorption profiles of Fig. 2b and six grid ranges. By choosing ideal materials for training and real materials for testing, it is ensured that the training and testing sets are significantly different.

Results are presented in the first column of Table 3. As can be seen, training with the entire dataset improves localization results compared to training using a single room and range (Table 2). Moreover, our method is able to estimate the mean wall absorption between 0 and 1 with an accuracy of 0.18, as well as the range of the source between 1 and 2.5m with an accuracy of 54cm. To the best of the authors' knowledge, the only other binaural range estimation method in the literature is [19], which showed an accuracy of about 1m using direct-to-reverberant ratios. Figure 3a shows mean source direction errors as a function of wall absorption. Note that too large or too little absorption hinders elevation estimation, the optimal results being obtained in the range $[0.3, 0.8]$. This is probably because a high-level of reverberation implies that model (1) is more approximate, while too little reverberation means less spatial richness, and less symmetry breaking. Similarly, Fig. 3b shows errors as a function of source distance. As commonly observed, farther sources are slightly harder to localize, although the method seems to be particularly robust to range variations.

Comparing now the first and second column of Table 3, we observe an interesting result. When removing diffusion in both training and test data, the estimation of all variables is degraded. It was consistently observed that adding diffusion in simulations improved results, even when using a number of other diffusion profiles and different room dimensions. This is particularly striking for elevation and absorption. We suspect diffusion to increase the spectral richness of binaural cues, making them more discriminative by breaking inherent symmetries of the problem. The authors are not aware of previous work specifically studying this effect in the literature.

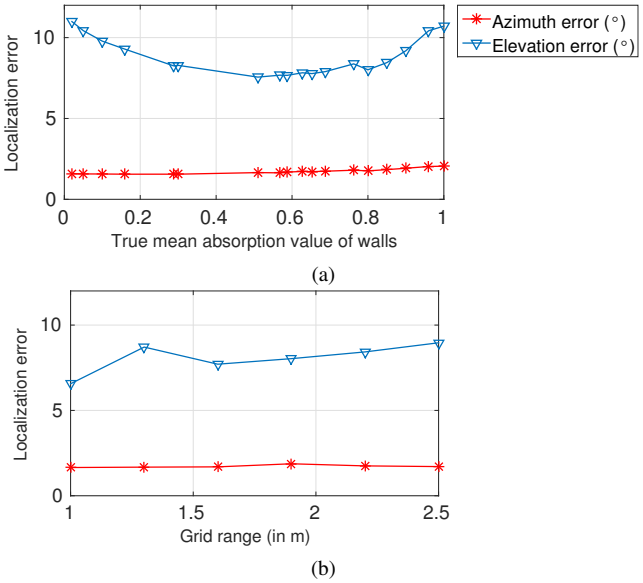


Fig. 3. Mean localization error (in degrees) as a function of absorption (a) and range (b).

We finally test the influence of removing absorption or range annotation during training (Columns 3 to 5 in Table 3). As expected, removing this information increases errors, but overall, the GLLiM probabilistic framework seems to be relatively robust to additional non-annotated effects such as absorption and range. This is promising for future experiments on larger datasets with larger parameter variability, where full annotation of all effects may not necessarily be possible.

5. CONCLUSION

In this paper, we presented a proof-of-concept for the novel framework of virtually-supervised learning for audio-scene geometry estimation. Obtained results are encouraging, revealing that estimating the 2D direction and range of a source as well as some of the wall acoustic properties is possible using binaural recordings only. Moreover, we observed that incorporating random diffusion effects in simulations significantly improved the spatial richness of binaural features, improving estimation of all parameters. In the future, larger training set variability (room size, receiver position) will be investigated, with the ultimate goal of testing the virtually-supervised method on real data. Extensions to speech [5], multiple sources [4] and additional partially-latent variables in GLLiM [18] will also be considered.

6. REFERENCES

[1] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[2] Nancy Bertin, S Kiti, and Rémi Gribonval, "Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6340–6344.

[3] Srđan Kitić, Laurent Albera, Nancy Bertin, and Rémi Gribonval, "Physics-driven inverse problems made tractable with cosparse regularization," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 335–348, 2016.

[4] Antoine Deleforge, Florence Forbes, and Radu Horaud, "Variational em for binaural sound-source separation and localization," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 76–80.

[5] Antoine Deleforge, Radu Horaud, Yoav Y Schechner, and Laurent Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 718–731, 2015.

[6] Bracha Laufer-Goldshtein, Ronen Talmon, and Sharon Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.

[7] Steven M Schimmel, Martin F Muller, and Norbert Dillier, "A fast and accurate "shoebox" room acoustics simulator," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 241–244.

[8] William M Hartmann, "Localization of sound in rooms," *The Journal of the Acoustical Society of America*, vol. 74, no. 5, pp. 1380–1391, 1983.

[9] Brad Rakerd and WM Hartmann, "Localization of sound in rooms, ii: The effects of a single reflecting surface," *The Journal of the Acoustical Society of America*, vol. 78, no. 2, pp. 524–533, 1985.

[10] Bill Gardner, Keith Martin, et al., "Hrft measurements of a kemar dummy-head microphone," 1994.

[11] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[12] Andrew Wabnitz, Nicolas Epain, Craig Jin, and André Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, 2010, pp. 1–6.

[13] Michael Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*, Springer Science & Business Media, 2007.

[14] Acoustic project company , Ukraine, "Absorption coefficients," [Online]. Available: http://www.acoustic.ua/st/web_absorption_data_eng.pdf, July 10, 2016.

[15] Michel Bruneau, *Fundamentals of acoustics*, John Wiley & Sons, 2013.

[16] Heinrich Kuttruff, *Room acoustics*, Crc Press, 2009.

[17] Adil Faiz, Joël Ducourneau, Adel Khanfir, and Jacques Chatillon, "Measurement of sound diffusion coefficients of scattering furnishing volumes present in workplaces," in *Acoustics 2012*, 2012.

[18] Antoine Deleforge, Florence Forbes, and Radu Horaud, "High-dimensional regression with gaussian mixtures and partially-latent response variables," *Statistics and Computing*, vol. 25, no. 5, pp. 893–911, 2015.

[19] Yan-Chen Lu and Martin Cooke, "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*, vol. 18, no. 7, pp. 1793–1805, 2010.