Mokslas – Lietuvos ateitis / Science – Future of Lithuania
ISSN 2029-2341 / eISSN 2029-2252
2020 Volume 12, Article ID: mla.2020.11462, 1–8
https://doi.org/10.3846/mla.2020.11462

Electronics and electrical engineering
Elektronika ir elektros inžinerija

DATASET FOR EVALUATION OF THE PERFORMANCE OF THE METHODS OF SOUND SOURCE LOCALIZATION ALGORITHMS USING TETRAHEDRAL MICROPHONE ARRAYS

Saulius SAKAVIČIUS*

Vilnius Gediminas Technical University, Vilnius, Lithuania

Received 03 November 2019; accepted 07 November 2019

Abstract. For the development and evaluation of a sound source localization and separation methods, a concise audio dataset with complete geometrical information about the room, the positions of the sound sources, and the array of microphones is needed. Computer simulation of such audio and geometrical data often relies on simplifications and are sufficiently accurate only for a specific set of conditions. It is generally desired to evaluate algorithms on real-world data. For a three-dimensional sound source localization or direction of arrival estimation, a non-coplanar microphone array is needed. Simplest and most general type of non-coplanar array is a tetrahedral array. There is a lack of openly accessible real-world audio datasets obtained using such arrays. We present an audio dataset for the evaluation of sound source localization algorithms, which involve tetrahedral microphone arrays. The dataset is complete with the geometrical information of the room, the positions of the sound sources and the microphone array. Array audio data was captured for two tetrahedral microphone arrays with different distances between microphones and one or two active sound sources. The dataset is suitable for speech recognition and direction-of-arrival estimation, as the signals used for sound sources were speech signals.

Keywords: audio dataset, sound source localization, room acoustics, tetrahedral microphone array, speech recognition, source separation.

Introduction

Sound source localization (SSL) and separation are some of the key elements in developing novel, speech-based human-machine interaction (HMI) systems. Information on sound source position in space or the direction-of-arrival (DoA) might be used to enhance audio and speech signals in such ambient intelligence systems, allowing for better source separation and thus higher quality of operation (Brutti et al., 2008). The development of methods and algorithms for sound source localization requires rigor testing on realistic data.

Most SSL algorithms rely on the usage of an array of microphones, signals from which are further processed to obtain an estimate of the direction of arrival (DoA) of the sound source, or the position of the source of the sound relative to the microphone array. The main classes of sound source DoA estimation are: a) time difference of arrival (TDoA) based methods; b) beamforming-based methods; and c) subspace transformation based methods (Lollmann et al., 2018).

It can be shown (Guentchev, 1997) that the minimum number of detectors required to obtain unambiguously a solution in three-dimensional space is four and that it is unique. Several authors researched into the localization of sound sources using tetrahedral microphone arrays (Alameda-Pineda & Horaud, 2014; Ozeki & Hamada, 2006). Nevertheless, authors did not release the audio data used for development and evaluation of their methods as an openly accessible dataset.

Recently, several learning-based sound source localization methods were proposed (Adavanne et al., 2017) (Takeda & Komatani, 2017; Chakrabarty & Habets, 2019). For learning-based SSL methods, a huge amount of training audio samples is needed. It is nearly impossible to produce such a large real-world audio dataset. Thus for such methods, a synthetic or semi-synthetic audio dataset is most often created (simulated). Nevertheless, it is desirable to evaluate the performance of such methods on real-world data. A concordance between the simulation and the real-world data is expected. Training audio and geometrical data can be simulated in a virtual environment,

*Corresponding author. E-mail: *saulius.sakavicius@vgtu.lt* which is modeled after a real-world counterpart. To achieve this, it is necessary to know exact parameters of the real-world environment, such as the dimensions and the acoustic properties of the room, the relative positions of the sound sources, and the microphones and the walls of the room. To be usable for the estimation of a three-dimensional sound source position or two-dimensional DoA (azimuth and elevation) estimation, the positions of the sound source in the dataset must not be coplanar and must exhibit at least some degree of variance in all three axes. For evaluation of SSL methods aimed at speech-based HMI systems, it is desirable that the signals of the sound sources in the dataset are human speech signals. While there are several audio datasets aimed at the SSL problems, they are all lacking some information or features described earlier: either the room dimensions or the position of the reference point relative to the room walls is unknown, or the sound sources are positioned on the same plane, or the signal of the sound sources is not speech. Thus, we present a simple dataset that satisfies all of the demands mentioned before: audio recordings are produced on a tetrahedral microphone array, using speech signals, with one or more than one simultaneously active sound source and with known dimensions of the room and the positions of the microphones and the sound sources relative to the walls of the room.

1. Previous work

There are several audio datasets presented earlier (Le Roux et al., 2015), focused on the sound source localization and separation tasks. The LOCATA dataset, presented as a part of IEEE-AASP Challenge on Acoustic Source Localization and Tracking, consists of audio recordings of one or two moving and up to four static sound sources, captured with a multitude of microphone arrays, with number of microphone per array ranging from 2 (binaural system using a dummy head) to 32 (*Eigenmike EM32* spherical array). The shortcoming of the LOCATA dataset is that neither the room dimensions nor the distance of the origin of the coordinate system to a corner of the room is presented, which imposes a limitation of usage of the LOCATA dataset for evaluation of learning-based SSL methods, such as presented by He et al. (2018, 2019) or Chakrabarty and Habets (2019), where the model is trained on semi-synthetic data, as it impossible to accurately simulate the environment matching the real-world. Also, the moving sound sources were the human subjects, walking in front of the microphone array and talking. Thus there is limited variance of the height of the sound sources relative to the origin of the coordinate system.

The Sound Source Localization for Robots (SSLR) Dataset is a collection of real robot audio recordings for the development and evaluation of sound source localization methods, recorded using Softbank robot Pepper, including robot ego-noise and overlapping multiple speech sources (We et al., 2018). The origin of the coordinate system for this dataset is the center of the microphone array, but no information is given about the room in which the dataset was collected nor the positions of the microphone array within those rooms. Moreover, the sound sources remain stationary, while the robot head is panning to sides. Thus the microphone-room spatial relationship is constantly changing, which is not the case in many ambient intelligence and surveillance systems, where the array is stationary for the duration of operation. Therefore, this dataset may not be well suited for evaluation of performance of static arrays.

Drone Egonoise and localization (DREGON) dataset is aimed at evaluating SSL using microphone arrays embedded in an unmanned aerial vehicle (UAV). The dataset contains both clean and noisy in-flight audio recordings continuously annotated with the 3D position of the target sound source using an accurate motion capture system (Strauss et al., 2018). The dataset includes the description of the room geometry and its reverberation time. Also, the speech signals were used for the static sound source. The downside of this dataset is that the microphone array is mounted on the UAV and is not stationary.

Collectively, none of the mentioned datasets feature a tetrahedral microphone array. We present a dataset for the evaluation of the performance of sound source localization algorithms that is captured by a static tetrahedral microphone array (two sets of experiments with different array geometries). We have used one or two static, simultaneously active sound sources with human speech signals. Our presented dataset includes thorough and explicit measurements of the room and the positions of the microphones and the sound sources with the origin of the coordinate system coinciding with one corner of the room.

2. Methods and materials

In this section, we present the methods for the dataset acquisition. For all audio recordings, a *Tascam US20×20* USB audio interface was used. All recordings were performed with a sampling rate f_s = 44.1 kHz and quantization resolution Q = 16 bit. All spatial measurements were made manually using a measuring tape with a precision of ±0.0005 m. The dataset consists of audio files of the microphone array, audio files of the sound sources, the room impulse response (RIR) measurement data, and the information about the positions of the sound sources, the microphones, and the geometry of the room. The array audio data was recorded for two array geometries. For each geometry, there were 10 cases of one active speech sound source and 10 cases for two active sound sources. As a result, a dataset of 40 different microphone and sound source combinations was produced, along with three RIR measurements, each using different combinations of source and microphone positions. The format and acquisition methods of each of these elements are discussed in the next section.

We also present the results of computer simulation using the image-source model for RIR generation, presented in (Allen & Berkley, 1976), of the same parameters as the real-world data to determine the level of discrepancies between the results of simulation and real-world RIR measurements.

2.1. Room properties

The dataset was acquired in a cuboid-shaped room in LinkMenų fabrikas, Vilnius Gediminas Technical University. The dimensions of the room were 5.400×5.860×2.640 m. The origin of the coordinate system of the dataset coincided with a corner of the room. Three of four of the walls of the room were made of painted masonry, while the fourth wall was a plaster wall. The volume of the room was $V = 89.869$ m^3 and the total surface area of the room was $S = 145.048$ m^2.

The furniture of the room consisted of three plywood tables, three chairs, several desktop computers, and computer monitors, which were not taken into account to not over-complicate the process of dataset acquisition.

The absorption coefficients of each of the wall were not directly measured but rather calculated from the measurement of the T_{60} reverberation time value using Sabine's equation (Sabine & Egan, 1994):

$$T_{60} = \frac{24\ln 10^1}{c_{20}}\frac{V}{Sa} \approx 0.1611\frac{V}{Sa}, \qquad (1)$$

Here c_{20} is the speed of sound at 20 °C and a is the average absorption coefficient of the surfaces of the room. Reordering (1) gives

$$a = 0.1611\frac{V}{ST_{60}}. \qquad (2)$$

The reverberation time can be calculated using Schroeder's method of backward integration of the RIR (Schroeder, 1965).

Schroeder's frequency F_c is calculated using an equation provided by Skålevik (2011):

$$F_c \approx 2000\left(\frac{T_{60}}{V}\right)^{0.5}. \qquad (3)$$

We have measured the impulse response of the room at three different combinations of the signal source, and the measurement microphone positions (microphone positions $M_{\mathrm{RIR},i}$, source positions $S_{\mathrm{RIR},i}$ and Euclidean distances between them $\Delta(M,S)_{\mathrm{RIR},i}$ are presented in Table 1 and in Figure 1).

For the RIR measurements, a *Mackie Thump12* powered loudspeaker was used as a sound source (axis of the loudspeaker directed to the capsule of the microphone). The measurement microphone was *Sonarworks XREF20*. RIR was captured using a MATLAB® tool *Room Impulse Measurer*. Provided by the tool are the two most widely used IR measurement techniques: Maximum-Length-Sequence (MLS) and Swept Sine. MLS technique is based on the excitation of the acoustical space by a periodic pseudo-random signal. The impulse response is obtained by calculating a circular cross-correlation between the measured output of the system and the excitation signal (Stan et al., 2002). The Swept Sine measurement technique uses an exponential time-growing frequency sweep as and the excitation signal. The output of the system is recorded, and deconvolution is used to recover the impulse-response from the swept sine tone (Farina, 2007). We have measured the impulse response using both techniques in all three source-microphone position combinations.

Table 1. Positions of the sound source and the microphone for the measurements of the RIR

	$S_{\mathrm{RIR},i}$			$M_{\mathrm{RIR},i}$			
i	x, m	y, m	z, m	x, m	y, m	z, m	$\Delta(M,S)_{\mathrm{RIR},i}$, m
1	4.160	3.740	1.030	1.395	0.730	1.520	4.116
2	4.490	2.430	0.295	3.825	1.145	1.490	1.877
3	4.980	1.400	1.030	4.765	3.275	0.340	2.009

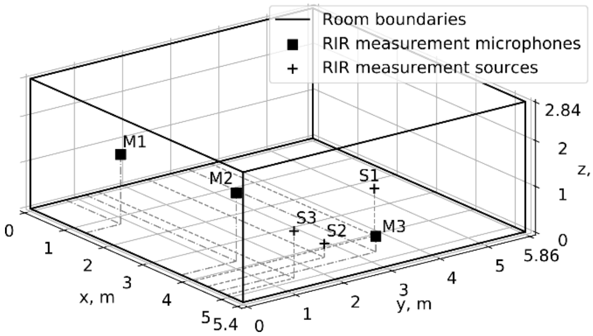


Figure 1. Positions of the RIR measurement microphones and sound sources within the room

2.2. Microphone arrays

We have obtained the audio recordings using two tetrahedral microphone arrays with different distances between the microphones (baseline length, B): ARRAY30 with B = 0.3 m and ARRAY60 with B = 0.6 m. This approach was chosen to allow the evaluation of the influence of the baseline length of the microphone array on the performance of the sound source localization algorithms. Maximum TDoA $\Delta T_{A_{\max}}$, observable using the array of baseline length B is

$$\Delta T_{A_{\max}} = \frac{B}{c_{20}}. \qquad (4)$$

For ARRAY30, $\Delta T_{A_{\max}30} = 8.82\times10^{-4}$ s. At $f_s = 44100$ Hz, this corresponds to 38 samples. For ARRAY60, the $\Delta T_{A_{\max}60} = 1.76\times10^{-3}$ s or 77 samples.

The positions of the microphones of both arrays are presented in Table 2 and Figure 2. The coordinates of the center of the array are calculated as the arithmetic mean of the coordinates of all microphones in each dimension.

Each tetrahedral array consists of four identical condenser microphones (*RØDE M2*). Since the directivity pattern of the *RØDE M2* microphone is cardioid shaped, we have positioned the microphones in such a way that the acoustic axes of the microphones were oriented upwards, so that the directivity of the microphones would be close to omnidirectional in a horizontal plane. The position reference point of each microphone coincided with the center of its membrane.

Table 2. Positions of microphones of ARRAY30 and ARRAY60

Array	*Mic No.*	*x, m*	*y, m*	*z, m*
ARRAY30	1	1.45	1.14	1.42
	2	1.425	0.84	1.42
	3	1.58	0.975	1.63
	4	1.295	1.025	1.63
	Center	*1.4375*	*0.995*	*1.525*
ARRAY60	1	1.49	1.325	1.36
	2	1.385	0.715	1.34
	3	1.72	0.975	1.78
	4	1.12	1.055	1.78
	Center	*1.42875*	*1.0175*	*1.565*

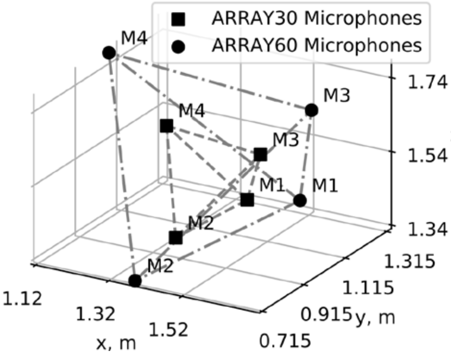


Figure 2. The positions of the ARRAY30 and ARRAY60 microphones; dashed lines denote the edges of the tetrahedrons

2.3. Sound sources

We have recorded the real-world audio data with each of the previously described array with one or two simultaneously active sound sources.

The sound sources were represented by two small loudspeakers: battery-powered *JBL GO* loudspeaker (Source 1, *SJ*), mounted on a tripod to allow for a convenient positioning; and *Yamaha MSP3* amplified two-way compact monitor loudspeaker (Source 2, *SY*), placed on a portable pedestal or a table.

The position of the sound source is determined by a reference point. For both sound sources the reference points were located in the center of the front grids of the speakers.

The speech signals that were reproduced through the speakers were obtained from the AMI Corpus (Carletta et al., 2006), headset microphone mix (file *ES2019a.Mix-Headset.wav*). To allow for the two simultaneously active sound sources to reproduce different signals, we have selected two excerpts from the file, each with a duration of 60 s. The first excerpt (E1) began at the 70-th second of the source audio file, and the second excerpt (E2) began at the 310-th second of the file.

Ten positions for Source 1 were randomly selected from a uniform distribution in the entire volume of the room. While all three coordinates were randomly chosen for the tripod-mounted Source 1, Source 2 could only be placed on a fixed height pedestal or the table. Thus its z coordinate z_2 is limited to two values: 0.85 m and 0.865 m above ground; x and y coordinates are the same for both source positions. The coordinates of the Source 1 (x, y, z_1) and Source 1 (x, y, z_2) of the selected positions are presented in Table 3. As can be seen from the Table 3, the average of coordinates of all source positions are very close to the geometric center of the room and differs from it no more than 8.25% (for x coordinate). The positions of the sources and the centers of both arrays are also presented in Figure 3.

By converting the Cartesian coordinates of the positions of the sound sources to polar coordinates, with the centers of the microphone arrays at the origin of the polar coordinate system, DoAs of sound sources were obtained (presented in Figure 4). DoA with azimuth $\theta = 0$ and elevation $\varphi = 0$ corresponds to the positive x axis of the Cartesian coordinate system.

Table 3. Selected positions for sound source placement in the room

Position No.	*x, m*	*y, m*	z_1*, m*	z_2*, m*
1	4	4.85	1.3	0.85
2	4.2	2.7	1.665	0.85
3	1.81	5.55	1.57	0.85
4	3.02	3.38	0.57	0.85
5	0.43	3.7	2.42	0.865
6	1.06	2.14	0.94	0.85
7	0.43	1.04	1.72	0.865
8	2.71	2.15	1.665	0.85
9	3.47	0.38	0.84	0.85
10	1.33	5.08	2.38	0.865
Standard deviation	1.423	1.734	0.613	0.007
Average	2.246	3.097	1.507	0.8545
Room center	2.69	2.925	1.42	1.42

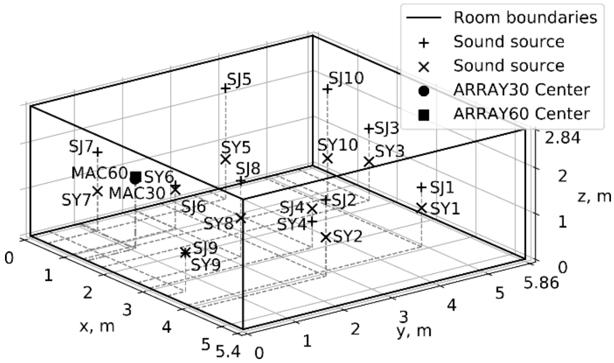


Figure 3. Positions of the sources (SJ$_i$ and SY$_i$ where i = 1, 2,..., 10 denotes the positions of Source 1 (*JBL GO*) and Source 2 (*Yamaha MSP3*) respectively, as presented in Table 3) and the centers of ARRAY30 (MAC301) and ARRAY60 (MAC601) within the room

For the single active sound source case, only Source 1 was used, and it was placed at all ten positions (coordinates of which are expressed as (x, y, z_1)). For the two active sound source case, ten positions of the Source 2 were selected from the Table 3 sequentially, while the positions of the Source 1 were selected from the Table 3 and randomly permuted, resulting in 10 combinations presented in Table 4. The speech signal excerpts were assigned to the sound sources in an alternating manner.

Table 4. Positions of sound sources in case of two simultaneously active sound sources and sources' corresponding signals

Source 2 position	Source 1 position	Source 2 signal	Source 1 signal
1	2	E2	E1
2	6	E1	E2
3	7	E2	E1
4	3	E1	E2
5	10	E2	E1
6	1	E1	E2
7	5	E2	E1
8	9	E1	E2
9	4	E2	E1
10	8	E1	E2

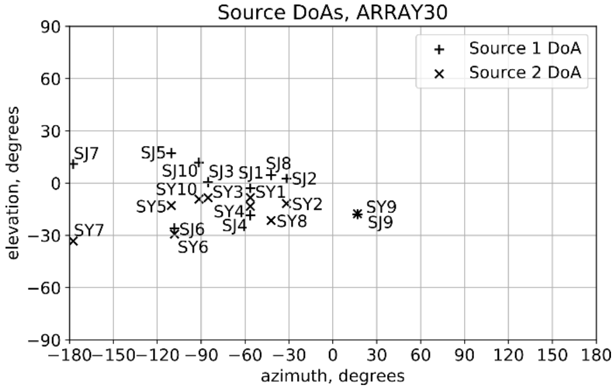


Figure 4. DoAs for source positions presented in Table 3, relative to the center of the ARRAY30

3. Results

To obtain the average absorption coefficient of the room *a*, a value of the T_{60} reverberation time is needed. This value was calculated from the impulse response of the room. The reverberation time T_{60} was calculated for each of the obtained RIR using Schroeder's backward integration method (Schroeder, 1965). The results are presented in Figure 5. The average T_{60} value was $\overline{T}_{60}$ = 552 ms, with standard deviation of 33.6 ms.

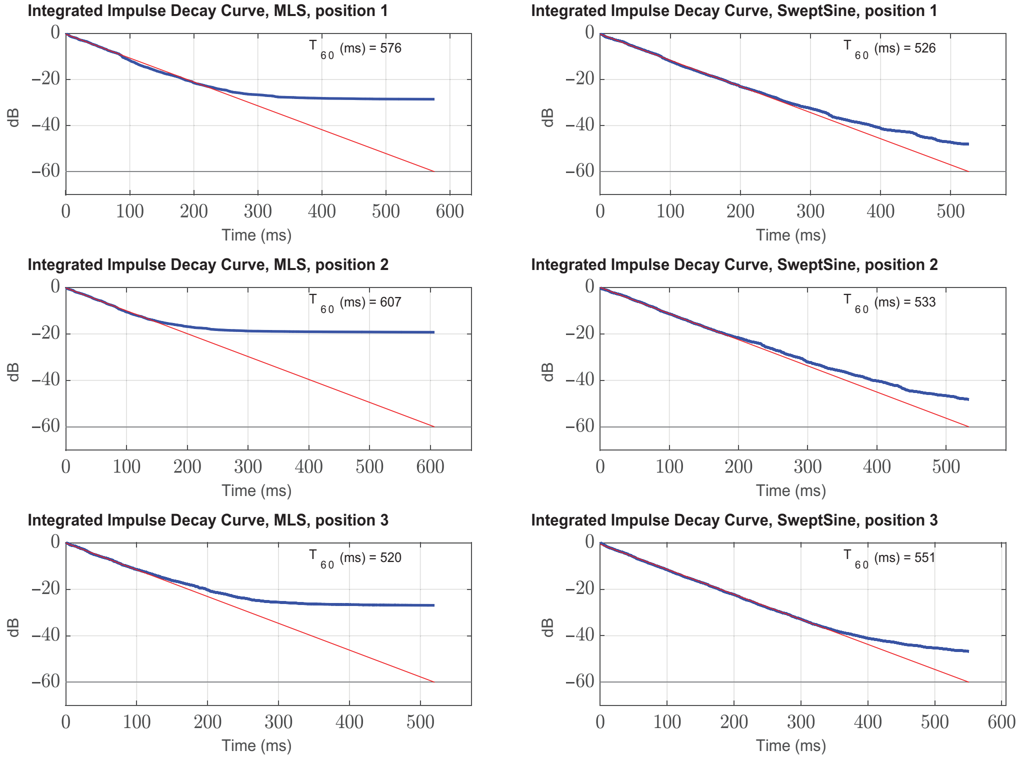


Figure 5. Results of the T_{60} estimation using Schroeder's backward impulse response integration (for all 3 RIR measurement source-microphone position combinations, using both MLS and Swept Sine RIR acquisition techniques)

The absorption coefficient was calculated using (2) with $T_{60} = 0.552$ s:

$$a = 0{,}1611 \frac{V}{S \cdot 0.552} = 0{,}206. \tag{5}$$

Schroeder's frequency was calculated using (3) and the measured room volume and T_{60}:

$$F_c = 2000 \left(\frac{0.552}{89.869} \right)^{0.5} = 156.76 \text{ Hz} \tag{6}$$

The measurements of RIRs were compared to the computer simulation of a virtual room with the same dimensions and the placement of the IR measurement sound source and microphone, using *Python* programming language and *pyroomacoustics* package, which uses image-source method for impulse response calculation (Scheibler et al., 2017). For the simulation, the absorption coefficient *a*, calculated in (5) was used, while the maximum order of reflection was 10. By performing the Fast Fourier Transform (FFT) of the RIRs, transfer functions of the room were obtained (magnitude spectra of the transfer functions presented in Figure 6).

As can be observed from the magnitude spectra of the transfer functions in all RIR measurement positions, the simulation is relatively accurate only in the approximate frequency range from 60 Hz to 500 Hz. This range starts at a frequency that is more than twice lower than Schroeder's frequency of the room and does not encompass the widely used telephone band (ITU-T, Rec. P.342, 2009). Thus, the auralization results using simulated RIRs might be inaccurate and unsuitable for reliable evaluation of the performance of sound source localization algorithms using speech signals. For all three measurement positions, the amplitude of the simulated transfer function is significantly higher in the low-frequency range than in measured RIRs. This can be addressed to a) unsuitability of the image source for RIR simulation in low frequency range (wave-based phenomena, such as diffraction and interference, are not properly recreated (Siltanen et al., 2010)) and b) inaccuracy of the real-world RIR measurements, as it relies on the linearity of the transfer functions of the transducers (measurement sound source and microphone, which are not linear. The diffraction effect is stronger at low frequencies where the wavelength is longer than or comparable to the dimensions of the reflecting objects (Siltanen et al., 2010), that is, lower than Schroeder's frequency. The frequency response of *Thump12* loudspeaker presents a steep roll-off in the sound pressure level below 70 Hz and above 6 kHz (Loud Technologies Inc., 2017), so it is impossible to obtain fully accurate RIR using neither Swept Sine nor MLS method using such loudspeaker. Considering these findings, it is advisable to evaluate SSL algorithms not only synthetic or semi-synthetic audio data but also on real-world audio data as the simulated audio signals might not accurately reflect the real-world situation.

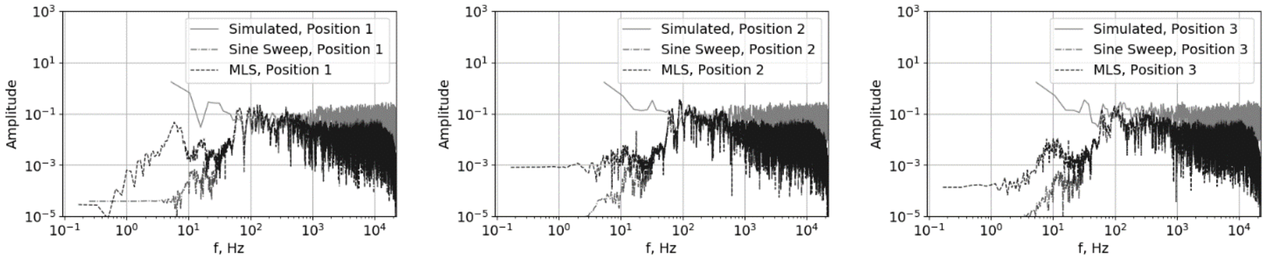
Our dataset is openly available online at https://github.com/Sakavicius/linkmenu-dataset.

Conclusions

A dataset of four different scenarios (two tetrahedral microphone arrays with different baseline lengths, one and two active sound sources for each type of array) was created, with ten different source positions (in case of two active sound sources – 10 two source position combinations) for each scenario. Positions of sound sources were distributed evenly in the room, with average of coordinates of all sources differing from the geometric center of the room no more than 8.25% (for *x* coordinate). A set of six room impulse responses was measured using three different combinations of source-microphone positions, using two IR acquisition techniques: MLS and Swept Sine. The reverberation time T_{60} was estimated from the RIR using Schroeder's method, and the average reverberation time $\overline{T}_{60}$ was determined to be 0.552 s. The average surface absorption coefficient was derived from the reverberation time and the geometry of the room and was determined to be *a* = 0.206. The Schroeder's frequency of the room was calculated to be 156.76 Hz.

A computer simulation of a virtual room with the same geometry and acoustical parameters as the real-world room was performed. From the comparison of results, it was determined that the magnitude spectra of real-world and simulated RIRs differ considerably both in low and high-frequency ranges, and the simulation is relatively accurate only in the approximate frequency range from 60 Hz to 500 Hz.

Thus, if a sound source localization method or algorithm is being developed, its evaluation of real-world audio is crucial as the simulated audio signals might not accurately reflect the real-world situation.



a) b) c)

Figure 6. Magnitude spectra of the transfer functions obtained from the RIR measurements (using Sine Sweep and MLS methods and computer simulation) at positions 1 (a), 2 (b) and 3 (c) (positions of sources and microphones presented in Table 1)

References

Adavanne, S., Politis, A., & Virtanen, T. (2017). Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. *ArXiv:1710.10059 [Cs, Eess]*. http://arxiv.org/abs/1710.10059

Alameda-Pineda, X., & Horaud, R. (2014). A geometric approach to sound source localization from time-delay estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(6), 1082–1095. https://doi.org/10.1109/TASLP.2014.2317989

Allen, J. B., & Berkley, D. A. (1976). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, *65*(4), 943–950. https://doi.org/10.1121/1.382599

Brutti, A., Omologo, M., & Svaizer, P. (2008). Localization of multiple speakers based on a two step acoustic map analysis. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4349–4352). https://doi.org/10.1109/ICASSP.2008.4518618

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kaldec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2006). The AMI meeting corpus: a pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction* (pp. 28–39). https://doi.org/10.1007/11677482_3

Chakrabarty, S., & Habets, E. A. P. (2019). Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, *13*(1), 8–21. https://doi.org/10.1109/JSTSP.2019.2901664

Farina, A. (2007, May). Advancements in impulse response measurements by sine sweeps. In *122nd Audio Engineering Society Convention* (pp. 2–21), Vienna, Austria.

Guentchev, K. (1997). *Learning-based three dimensional sound localization using a compact non-coplanar array of microphones* (Master's thesis). Department of Computer Science, Michigan State University.

He, W., Motlicek, P., & Odobez, J.-M. (2018). Deep neural networks for multiple speaker detection and localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 74–79). https://doi.org/10.1109/ICRA.2018.8461267

He, W., Motlicek, P., & Odobez, J.-M. (2019). Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 770–774). https://doi.org/10.1109/ICASSP.2019.8682655

ITU-T, Rec. P.342. (2009). *Transmission characteristics for narrow-band digital loudspeaking and hands-free telephony terminals*. International Telecommunication Union, Geneva.

Le Roux, J., Vincent, E., Hershey, J. R., & Ellis, D. P. W. (2015). Micbots: Collecting large realistic datasets for speech and audio research using mobile robots. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5635–5639). https://doi.org/10.1109/ICASSP.2015.7179050

Lollmann, H. W., Evers, C., Schmidt, A., Mellmann, H., Barfuss, H., Naylor, P. A., & Kellermann, W. (2018). The LOCATA challenge data corpus for acoustic source localization and tracking. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)* (pp. 410–414). https://doi.org/10.1109/SAM.2018.8448644

Loud Technologies Inc. (2017). *Thump12A, Thump15A 1300W powered loudspeakers. Owner's manual.* https://mackie.com/sites/default/files/PRODUCT%20RESOURCES/MANUALS/Owners_Manuals/Thump12A_Thump15A_OM.pdf

Ozeki, K., & Hamada, N. (2006). Estimating directions of multiple sound sources using tetrahedral microphone array. In *TENCON 2006 – 2006 IEEE Region 10 Conference* (pp. 1–4). https://doi.org/10.1109/TENCON.2006.343853

Sabine, W. C., & Egan, M. D. (1994). *Collected papers on acoustics*. Harvard University Press. https://doi.org/10.1121/1.409944

Scheibler, R., Bezzam, E., & Dokmanić, I. (2017). Pyroomacoustics: A Python package for audio room simulations and array processing algorithms. *ArXiv:1710.04196 [Cs, Eess]*. http://arxiv.org/abs/1710.04196

Schroeder, M. R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, *37*(3), 409–412. https://doi.org/10.1121/1.1909343

Siltanen, S., Lokki, T., & Savioja, L. (2010). Rays or waves? Understanding the strengths and weaknesses of computational room acoustics modeling techniques. In *Proceedings of the International Symposium on Room Acoustics, ISRA 2010* (pp. 1–6).

Skålevik, M. (2011). *Schroeder frequency revisited.* Paper presented at the Proceedings of Forum Acusticum.

Stan, G. B., Embrechts, J. J., & Archambeau, D. (2002). Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, *50*(4), 249–262.

Strauss, M., Mordel, P., Miguet, V., & Deleforge, A. (2018). DREGON: dataset and methods for UAV-embedded sound source localization. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1–8). https://doi.org/10.1109/IROS.2018.8593581

Takeda, R., & Komatani, K. (2017). Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2217–2221). https://doi.org/10.1109/ICASSP.2017.7952550

DUOMENŲ RINKINYS GARSO ŠALTINIO LOKALIZAVIMO, TAIKANT TETRAEDRINES MIKROFONŲ GARDELES, METODŲ CHARAKTERISTIKOMS TIRTI

S. Sakavičius

Santrauka

Garso šaltinio lokalizavimo ir išskyrimo algoritmams kurti ir charakteristikoms tirti reikalingas nuosekliai sudarytas garso duomenų rinkinys, papildytas informacija apie akustines patalpos savybes, garso šaltinių ir mikrofonų gardelės padėtis. Dažnai tokie garso ir geometriniai duomenys gaunami atliekant kompiuterinę emuliaciją, tačiau dauguma emuliacijos metodų grindžiami supaprastinimais ir yra tikslūs tik tam tikromis sąlygomis. Todėl garso šaltinio lokalizavimo ir išskyrimo algoritmų veikimą išsamiai įvertinti galima tik taikant realius garso duomenis. Siekiant nustatyti garso šaltinio padėtį ar sklidimo kryptį erdvėje, reikalinga mikrofonų gardelė, kurios elementai yra nekomplanarūs. Paprasčiausias ir bendriausias nekomplanarios gardelės tipas yra tetraedrinė gardelė. Šiuo metu nėra laisvai prieinamo garso ir geometrinių duomenų rinkinio, surinkto naudojant tokio tipo mikrofonų gardeles. Šiame straipsnyje pristatomas duomenų rinkinys, skirtas garso šaltinio lokalizavimo ir išskyrimo algoritmams tirti naudojant tetraedrines mikrofonų gardeles. Duomenų rinkinį sudaro garso duomenys ir juos atitinkanti geometrinė informacija: patalpos matmenys, garso šaltinių ir mikrofonų gardelės padėtys patalpos atžvilgiu. Garso duomenys buvo surinkti naudojant dvi tetraedrines mikrofonų gardeles su skirtingais atstumais tarp mikrofonų, esant vienam arba dviem vienu metu aktyviems garso šaltiniams. Garso šaltiniais buvo atkuriamas žmogaus kalbos signalas, todėl pristatomas duomenų rinkinys yra tinkamas kalbos atpažinimo ir sklidimo krypties nustatymo algoritmams tirti.

Reikšminiai žodžiai: garso duomenų rinkinys, garso šaltinio lokalizavimas, patalpos akustika, tetraedrinė mikrofonų gardelė, kalbos atpažinimas, garso šaltinio išskyrimas.