THE SINGLE- AND MULTICHANNEL AUDIO RECORDINGS DATABASE (SMARD)

Jesper Kjær Nielsen[†], Jesper Rindom Jensen[‡], Søren Holdt Jensen[†], and Mads Græsbøll Christensen[‡]

[†]Aalborg University
Dept. of Electronic Systems
{jkn,shj}@es.aau.dk

[‡]Aalborg University
Audio Analysis Lab, AD:MT
{jrj,mgc}@create.aau.dk

ABSTRACT

A new single- and multichannel audio recordings database (SMARD) is presented in this paper. The database contains recordings from a box-shaped listening room for various loudspeaker and array types. The recordings were made for 48 different configurations of three different loudspeakers and four different microphone arrays. In each configuration, 20 different audio segments were played and recorded ranging from simple artificial sounds to polyphonic music. SMARD can be used for testing algorithms developed for numerous application, and we give examples of source localisation results.

Index Terms— Multichannel recordings, audio database, source localisation

1. INTRODUCTION

The processing of single- and multichannel audio recordings is a very active field of research within the signal processing community and important in numerous applications such as noise reduction, echo cancellation, source separation, source localisation, room geometry estimation, distributed array processing, recognition, and de-reverberation [1–11]. Despite the heavy research activities in the field, only a few high quality multichannel audio recordings have been made available online to the research community. Consequently, many of the developed signal processing algorithms are in research papers only evaluated on simulated data (e.g., generated using room impulse response generators [12, 13]), on data of a low quality, or on non-public data, and this might inhibit reproducibility [14], complicate algorithm comparison, or ultimately lead to the wrong conclusions. There are multiple reasons for why it is difficult to obtain good quality audio recordings. For example, making high quality recordings can be very time-consuming; professional measurement equipment is often expensive; dedicated listening rooms, anechoic chambers, and laboratory equipment might not be available; or many external nuisances, which might be hard to eliminate, can influence the quality of the recordings.

This work was partially funded by the Danish Council for Independent Research - Grant no.: DFF–1337-00084, the Villum Foundation, and InnovationsFonden.

To extend the amount of freely available single- and multichannel audio recordings, we here present a new database called the *Single- and Multichannel Audio Recordings Database* (SMARD) which is made freely available online[1]. The database contains both single- and multichannel recordings of artificial signals as well as of reverberant and anechoic speech, vocal, and musical signals emitted by various loudspeaker types. The recordings are made with four different microphone arrays in a box-shaped listening room. Moreover, the position of the microphones and loudspeakers inside this room as well as the temperature are also measured. Although useful in many applications, SMARD was compiled for the primary purpose of source localisation and room geometry estimation. For source localisation, SMARD extends databases such as the multi-channel Wall Street journal audio visual corpus (MC-WSJ-AV) [15] and the audio-visual corpus for speaker localization and tracking [16] by using multiple array topologies and other source signals than speech. Moreover, since the microphone and loudspeaker locations relative to the room were also measured, SMARD can also be used for room geometry estimation. Although the data used in [4] have been made available online, SMARD is to the best of our knowledge the first freely available database which can be used for this purpose. For various array structures and settings, estimated multichannel impulse responses can be found in online databases[2]. In contrast to these, SMARD also contains the raw recordings for a large number of source signals.

In this paper, we first describe how the data in SMARD were recorded. Specifically, we describe the listening room, the measurement equipment, and the measurement configurations. Secondly, we give a few examples of source localisation results based on these data.

2. DESCRIPTION OF SMARD

SMARD contains approximately 18 GB of audio recordings at a sampling frequency of 48 kHz. All or a subset of these

[1]The data can be downloaded at http://www.smard.es.aau.dk (DOI: 10.5278/vbn/misc/smard). If you use SMARD in a peer-reviewed research paper or in a book, we kindly ask you to email us a BibTeX entry which we will publish on the SMARD website.

[2]See http://www.commsp.ee.ic.ac.uk/ for a list of these databases.

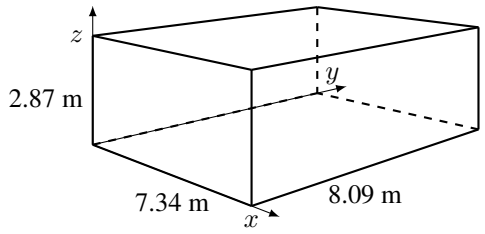


Fig. 1: A sketch of the multi-channel listening room.

Loudspeakers	Brüel & Kjær OmniPower 4296 Brüel & Kjær OmniSource 4295 Custom-made 3" directional loudspeaker
Microphones	G.R.A.S. Prepolarized Free Field Microphone 40AZ G.R.A.S. 26CC Microphone pre-amplifier (dummy) Bruel & Kjær JJ-2617 Coaxial input adapter (51 pF)
Arrays	Single microphone Three 7-element ULAs with 5 cm spacing Two 6-element UCAs with radii of 4 cm and 6 cm Orthogonal array consisting of the three ULAs
A/D Converters Sound Card Power amplifier	Behringer Ultragain Pro-8 Digital ADA8000 RME Digi 9652 Project Hammerfall Rotel RB-976
Software	Playrec 2.1.0 Portaudio v. 19 20071207 Steinberg ASIO SDK 2.3 MATLAB 2013b

Table 1: Measurement Equipment. More details can be found on the SMARD website.

recordings as well as estimated impulse responses can be downloaded from the SMARD website.

2.1. The Room

The recordings have all been made in a 60 m^2 multichannel listening room at Aalborg University, and a sketch of this room is given in Fig. 1. The room is box-shaped, symmetrical, and has been measured using the Brüel & Kjær Type 2270 to have a reverberation time of approximately 0.15 seconds.

2.2. Equipment

Recordings were made for various combinations of different loudspeakers and microphone arrays. As detailed in Table 1, three loudspeakers were used. The Brüel & Kjær OmniPower 4296 and the Brüel & Kjær OmniSource 4295 are both approximately omnidirectional loudspeakers within a limited frequency range. The OmniPower 4296 loudspeaker can emit more sound power than the OmniSource 4295, but can only be considered omnidirectional over a narrower frequency range. The directional loudspeaker is a conventional 3" speaker in a wooden cabinet. Pictures of it and its on-axis impulse response can be found on the SMARD website.

Up to 22 G.R.A.S. microphones were used in various array configurations. The simplest array was just a single microphone which can been seen on the right hand side of Fig. 2a. The other array types were uniform linear arrays (ULA) (see Fig. 2c), uniform circular arrays (UCAs) (see Fig. 2d), and an orthogonal array (see Fig. 2b). Finally, a *dummy* microphone, which is simply a capacitor mounted on a microphone pre-amplifier, was also present in all recordings. The recordings from the dummy microphone can be used to inspect electrical noise, cross-talk, etc. Except for the loudspeakers, microphones, and arrays, all measurement equipment was situated in a control room adjacent to the multichannel listening room. The equipment is listed in Table 1.

Loudspeaker type	
0XXX	OmniPower 4296
1XXX	OmniPower 4295
2XXX	Directional loudspeaker
Loudspeaker position and orientation	
X0XX	Placed at $(2.00, 6.50, 1.25)$. Angle of -90° in XY-plane
X1XX	Placed at $(3.50, 4.50, 1.50)$. Angle of -45° in XY-plane
Array types	
XX0X	Orthogonal array, single microphone, and dummy microphone
XX1X	Three ULAs and dummy microphone
XX2X	Two UCAs, one ULA, and dummy microphone
Array positions and orientations	
XX00	See the SMARD website
XX01	See the SMARD website
XX02	See the SMARD website
XX03	See the SMARD website
XX10	See the SMARD website
XX11	See the SMARD website
XX20	See the SMARD website
XX21	See the SMARD website

Table 2: The 48 measurement configurations.

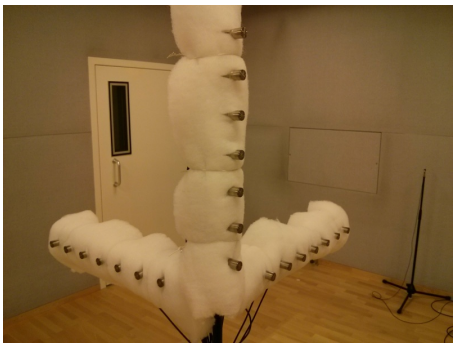
Artificial Sounds	
1	Five seconds of silence
2	Exponential sine sweep (ESS) from 10 Hz to 24 kHz
3	Harmonic signals with increasing fundamental frequency in steps
4	Eight repetitions of a 16th order MLS sequence
5	Pink noise
6	Single sinusoidal tone with increasing frequency in steps
7	White Gaussian noise
Speech/vocal signals	
8	Soprano vocal from the EBU SQAM CD
9	Quartet vocal from the EBU SQAM CD
10	Male voice from the EBU SQAM CD
11	Child's voice from the TSP speech database
12	Female voice from the TSP speech database
13	Male voice from the TSP speech database
Musical signals	
14	Clarinet from the EBU SQAM CD
15	Trumpet from the EBU SQAM CD
16	Xylophone from the EBU SQAM CD
17	Abba excerpt from the EBU SQAM CD
18	Bass flute from the MIS database
19	Guitar from the MIS database
20	Violin from the MIS database

Table 3: The 20 audio segments.

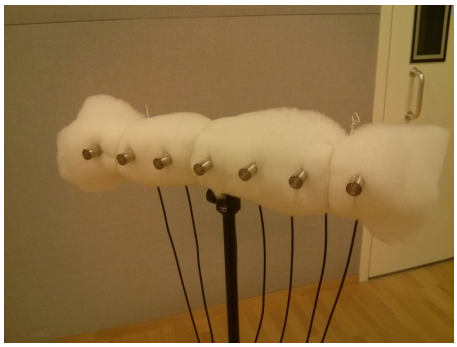
2.3. Measurements Configurations

Measurements were made for a total of 48 different configurations. Each configuration is enumerated by a four digit number of the form ABCD where the first and most significant digit A denotes the type of loudspeaker; the second digit B denotes the position and orientation of the loudspeaker; the

(a) Configuration number 0002.



(b) The orthogonal array.



(c) A uniform linear array.

(d) The uniform circular arrays.

Fig. 2: Some pictures of the measurement setup and the arrays.

third digit C denotes the type(s) of microphone arrays; and the least significant digit D denotes the position and orientations of these arrays. Table 2 summarises these configurations and further details can be found on the SMARD website. As an example, Fig. 2a shows a picture of configuration number 0002 which include the OmniPower 4296, the orthogonal array, the single microphone, and the dummy microphone.

2.4. Audio Segments

For each of the 48 configurations, a total of 20 audio segments were played and recorded. As listed in Table 3, seven artificial signals, six speech/vocal signals, and seven musical signals were used. The artificial signals were all created in MATLAB and a description of how they were generated can be found on the SMARD website. The speech and musical signals consist of both reverberant and anechoic signals. The signals from the EBU SQAM CD [17] are reverberant signals whereas the signals from the TSP speech database [18] and the musical instrument samples (MIS) database [19] are anechoic signals. Reverberant signals were included in the set of audio segments since the localisation of loudspeakers often involves such source signals. All of these databases are freely available online[3][4][5] for research usage.

[3]EBU SQAM CD: https://tech.ebu.ch/publications/sqamcd

[4]TSP speech database: http://www-mmsp.ece.mcgill.ca/Documents/Downloads/TSPspeech/

[5]MIS database: http://theremin.music.uiowa.edu/MIS.html

For every configuration, the temperature inside the multichannel listening room was measured and stored before these 20 audio segments were played. For each of the audio segments, all of the microphone recordings and a loop-back of the loudspeaker signal were stored in the database. A pause of two seconds was added between the segments to ensure that the sound field within the room was stationary before the next segment was played. The first audio segment was just five seconds of silence. The recordings made with this input signal can be used to inspect the stationary acoustical background noise. From the audio recordings containing the ESS, we have also estimated the room impulse responses (RIR) from the speaker to the different microphones. These, and further details about how they were estimated, are available from the download section of the website.

3. EXAMPLES OF USE

As previously mentioned, SMARD is useful for evaluating, e.g., noise reduction, localisation, and room geometry estimation algorithms. In this section, we present some results obtained from the evaluation of two localisation algorithms on some of the data in SMARD. The evaluated algorithms are the steered response power with phase transform (SRP-PHAT) method [20], and a near-field, maximum likelihood (ML) method recently proposed in [21]. As the ML method assumes that the desired signal is quasi-periodic, the methods

		Synthetic						Violin					
Config. no.		2000	2001	2002	2003	2020	2021	2000	2001	2002	2003	2020	2021
True	φ	111.6	111.1	130.2	139.4	108.4	134.8	111.6	111.1	130.2	139.4	108.4	134.8
	ψ	3.1	-2.3	-3.2	2.8	-3.9	2.9	3.1	-2.3	-3.2	2.8	-3.9	2.9
	r_c	5.3	5.3	3.8	5.9	4.6	4.8	5.3	5.3	3.8	5.9	4.6	4.8
SRP-P	φ	112.7	111.8	131.3	141.1	99.1	130.8	112.7	111.1	131.4	140.4	112.5	147.1
	ψ	-0.7	-27.2	-4.4	-22.1	-15.1	15.0	2.6	-1.7	-7.6	-0.4	3.7	4.2
	r_c	4.4	2.8	5.6	2.4	4.3	2.9	6.7	6.7	6.0	7.9	2.8	3.9
ML-AP	φ	111.7	112.1	131.2	140.3	112.8	133.7	112.4	110.9	131.4	139.9	105.8	138.1
	ψ	-0.4	-26.1	-4.8	-20.3	-12.2	14.0	4.3	-1.5	-5.2	0.5	1.7	2.5
	r_c	7.0	3.2	5.2	2.6	6.0	1.6	5.6	3.0	6.1	8.3	6.1	0.4

Table 4: Location estimates in spherical coordinates for different configurations.

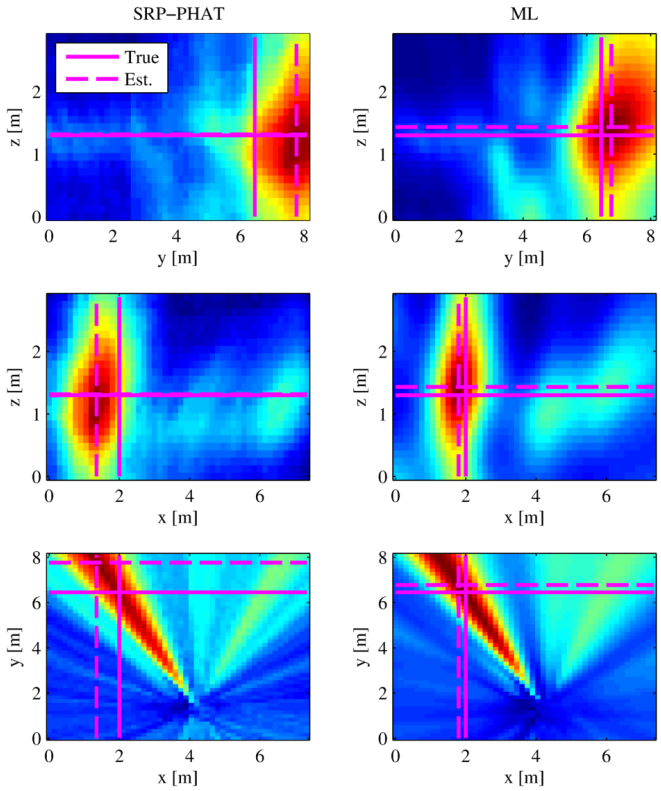


Fig. 3: Cost-functions versus cartesian coordinates for the SRP-PHAT and ML methods when applied on configuration 2000.

were applied on the synthetic harmonic signals and the violin signals. More specifically, a single segment of 100 samples was used from each microphone in different configurations. The segment from the harmonic signals was taken from the last part of the signal where the pitch is 500 Hz, while the segment from the violin signals was taken from the first part. The pitch of the signals, which is needed in the ML method, is estimated using the method in [22], and the number of harmonics was assumed known. Further details about the simulation setup can be found on the SMARD website along with the code for running the simulations.

With this simulation setup, we first of all obtained the results in Figure 3, depicting cost functions and location estimates for the SRP-PHAT and ML methods when applied for localisation of the violin source in configuration 2000. To obtain these plots of the cost functions versus two coordinates at a time, the last coordinate was fixed to the value estimated by the method. We see from the results that the cost functions peak relatively close to the true source position. The methods were also evaluated on other scenarios, yielding the results in Table 4. Generally, the angle (azimuth φ and elevation ψ) estimates are close to the true angles except for a few cases where a strong reflection from the wooden floor dominates the cost functions (configurations 2001 and 2003). The range (r_c) estimates are more inaccurate, but in most cases, the source is also placed relatively far from the arrays. These results supports the potential of applying SMARD for evaluation of, e.g., localisation methods, and the validity of the recorded data.

4. CONCLUSIONS

We have here presented the Single- and Multichannel Audio Recordings Database (SMARD) which can be used for testing algorithms developed for numerous audio signal processing tasks such as source localisation and room geometry estimation. SMARD is made freely available online to facilitate easier testing on real recordings, reproducibility of results, and algorithm comparison based on the same data. The database contains multichannel recordings for 20 audio segments in 48 different configurations arising from using three different loudspeakers, four different microphone arrays, and various source and sensor locations inside a box-shaped listening room.

Acknowledgements

The authors would like to thank Claus Vestergaard Skipper, Martin Bo Møller, Neo Kaplanis, and Peter Skotte for assisting in solving practical problems w.r.t. making the measurements.

5. REFERENCES

[1] F. Antonacci, J. Filos, M. Thomas, E. Habets, A. Sarti, P. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.

[2] J. Benesty, J. Chen, and Y. A. Huang, *Microphone array signal processing*, Berlin, Germany: Springer-Verlag, 2008.

[3] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, New York, NY, USA: Springer-Verlag, 2001.

[4] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 30, pp. 1–6, 2013.

[5] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Signals and Communication Technology. Springer, 2010.

[6] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Institute of Electrical and Electronics Engineers, 2000.

[7] F. Küch and W. Kellermann, *Nonlinear Acoustic Echo Cancellation*, Springer, Heidelberg, Germany, 2006.

[8] S. Makino, T. W. Lee, and H. Sawada, *Blind Speech Separation*, Signals and Communication Technology. Springer, 2007.

[9] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, Springer-Verlag, 2008.

[10] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. Symp. Commun., and Veh. Technol.*, Nov. 2011, pp. 1–6.

[11] R. Heusdens, G. Zhang, R. C. Hendriks, Y. Zeng, and W. B. Kleijn, "Distributed MVDR beamforming for (wireless) microphone networks using message passing," in *Proc. Intl. Workshop Acoust. Echo Noise Control*, Sep. 2012, pp. 1–4.

[12] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010, Ver. 2.0.20100920.

[13] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 269–277, Jul. 2008.

[14] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 37–47, May 2009.

[15] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE Workshop on Autom. Speech Recog. and Underst.* IEEE, 2005, pp. 357–362.

[16] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking," in *Proc. Int. Workshop on Machine Learning for Multimodal Interaction*. 2005, pp. 182–195, Springer.

[17] European Broadcasting Union, "Sound quality assessment material recordings for subjective tests: Users' handbook for the EBU SQAM CD," Tech. Rep. EBU – TECH 3253, European Broadcasting Union, 2008.

[18] P. Kabal, "TSP speech database," Tech. Rep., Dept. of Electrical & Computer Engineering, McGill University, 2002.

[19] L. Fritts, "University of Iowa Musical Instrument Samples," http://theremin.music.uiowa.edu/, 1997.

[20] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., chapter 8, pp. 157–180. Springer-Verlag, 2001.

[21] J. R. Jensen and M. G. Christensen, "Near-field localization of audio: a maximum likelihood approach," in *Proc. European Signal Processing Conf.*, Sep. 2014, submitted.

[22] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "Default Bayesian estimation of the fundamental frequency," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 598–610, Mar. 2013.