2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914724

THE SEBASS-DB: A CONSOLIDATED PUBLIC DATA BASE OF LISTENING TEST RESULTS FOR PERCEPTUAL EVALUATION OF BSS QUALITY MEASURES

Thorsten Kastner, and Jürgen Herre

International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany
{thorsten.kastner,juergen.herre}@audiolabs-erlangen.de

ABSTRACT

For the development of new and improvement of existing perceptual quality measurement methods for Blind Audio Source Separation (BASS), a large and comprehensive body of subjective reference ratings is necessary, originating from well-conducted listening tests, that enable calibration and testing. To support the community in the advancement of perception-based quality measures for BASS, we published the SEBASS-DB , i.e. the Subjective Evaluation of Blind Audio Source Separation Data Base. The SEBASS-DB is a meta dataset containing the results from five high-quality MUSHRA based listening tests on the Basic Audio Quality of (blindly) separated audio source signals and can be freely used for non-commercial purposes. It contains a total of over 11,000 ratings from over 900 rated audio signals submitted to the audio source separation campaigns 2007, 2008 and 2018 as well as listening test results from three other publications by the authors. Major parts of the listening tests were conducted by expert listeners.

Index Terms— audio source separation, listening tests, dataset

1. INTRODUCTION

Recently, Blind Audio Source Separation (BASS) has been receiving more and more attention due to increasing use of deep learning based methods and due to its manifold application scenarios like speech or dialogue separation / enhancement, noise suppression or personal re-mixing of audio material. In most of these applications, BASS-generated signals are listened to by a human subject and thus their perceived quality is a crucial criterion for user acceptance and satisfaction. Using properly conducted listening tests for perceptual quality assessment of BASS technology can, however, be time and resource consuming. In contrast, perception-based computational methods to automatically grade the perceived quality of BASS systems offer an efficient and reproduceable alternative. Thus, the availability of good perception-based quality measurement methods for BASS is a key factor in the advancement of BASS method. A crucial factor is the available amount of perception-based computational measures of reference data for calibrating and testing. While it is often a simple matter to generate sufficient reference data for testing purely signal-driven, energy-based measures, subjective reference ratings must be collected for the development of perception-based measures. However, properly conducted listening tests under controlled conditions with experienced listeners are time and resource consuming and thus their results are welcomed in the community. Since there can never be enough of such data available for training, testing or validating, we released the SEBASS-DB [1] and made it available to the community for free for research purposes but described it only briefly in [8] in 2019.

The SEBASS-DB is a meta dataset of high quality MUSHRA-based listening tests conducted in the course of three research projects [6–8] on the evaluation and development of perception-based measures for BASS systems. Expert listeners took part in most of the tests and rated the Basic Audio Quality (BAQ) [2] of (blindly) separated audio source signals. BAQ was chosen as a quality criterion as it defines a general quality criterion for evaluating the perceived quality of an audio signal and has also been proposed as a general audio quality criterion for subjective evaluation of audio source separation algorithms [9].

The SEBASS-DB contains ratings from separated source signals from the Stereo Audio Source Separation Evaluation Campaign 2007 (SASSEC) [3], from the signal Separation Evaluation Campaign 2008 (SiSEC08) [4] and from the SiSEC18 MUSDB18 dataset [10]. Further listening test results from separated source signals from the PEASS dataset [11] re-rated regarding BAQ and listening test results from a study on the influence of the rendering architecture on the subjective performance of source separation algorithms [7]. The following individual datasets of listening test results are included in the SEBASS-DB:

- SASSEC dataset
- SiSEC08 dataset
- SAOC DB dataset
- PEASS BAQ dataset
- SiSEC18 dataset

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 25,2024 at 14:00:18 UTC from IEEE Xplore. Restrictions apply.

All datasets can be downloaded from [1].

The remainder of the paper is organized as follows. Section 2 describes the common listening test procedure, i.e. how the items were presented, how the responses were collected and how the tests were conducted. The SEBASS-DB itself is described in detail in Section 3. A textual summary is presented for each of the five datasets including a overview table of statistical key figures at the end of the section. Finally, Section 4 presents concluding remarks.

2. DESCRIPTION OF LISTENING TEST EFFORT

All listening tests were performed in the same way. Test and reference signals were presented in the same way and the listener's response to the stimuli was quantified in the same manner.

2.1. Stimulus Presentation

All listening tests were based on recommendation ITU-R BS.1534-3 [2] (nicknamed "MUSHRA") for assessing intermediate audio quality and carried out in accordance with guidelines for subjective evaluation of audio source separation algorithms proposed by Vincent et al. [9]. In each trial the participants had to blindly rate separated versions of the source signal, the original source signal (hidden reference) and at least one anchor signal in comparison to the known original source signal (reference). Unlike Recommendation ITU-R BS.1534-3, the signal mixture was always used as an anchor signal instead of a 3.5 kHz low-pass filtered version of the reference signal. The participants could switch instantaneously between each of the presented signals and had the possibility to set playback loops. The listening tests were conducted in the sound laboratories of the Fraunhofer Institute IIS and the International Audio Laboratories Erlangen. The items were presented via headphones. Binaural rendering was not applied. Within a listening test the audio signals had the same duration, but between different listening tests the duration varied between 5 to 10 seconds. The participants had the possibility to set the playback volume at the beginning of the listening test.

2.2. Listener Response Collection

In each trial of the listening test the participants graded the unknown signals (hidden reference, anchor signal(s) and signals under test) on the MUSHRA scale (0 to 100) in comparison to the known original source signal (reference). The numerical scale is equally divided into five segments and semantically annotated ("bad", "poor", "fair", "good", "excellent"). The hidden reference earns a grade of 100. The question asked to the participants was: "Grade the Basic Audio Quality of the items under test with respect to the reference signal. All detected differences must be interpreted as impairment.".

2.3. Conduction of Listening Tests

The listening tests were usually divided into several sessions such that one session did not last longer than half an hour to avoid listener fatigue. A training session was obligatory containing approximately 10% of the items of the listening test. An information sheet was prepared describing the procedure of the listening test and the utilized MUSHRA program. webMUSHRA [12] (Fig. 1), a web browser based application with graphical user interface for conducting listening tests, was used for the SiSEC18 listening tests. An internal application with similar user interface and guidance was used for all other tests.

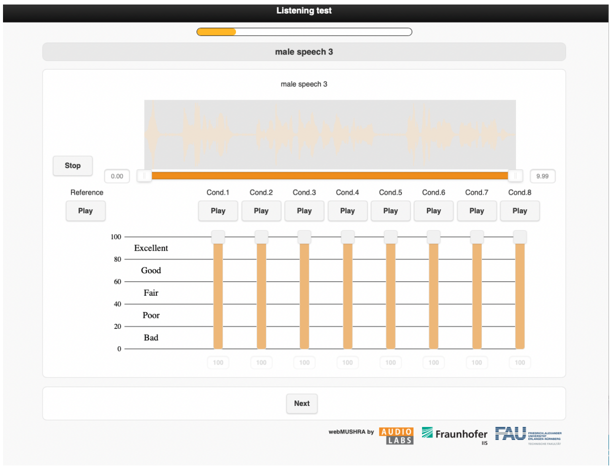

Fig. 1. Graphical user interface of the webMUSHRA software

3. THE SEBASS DB

The SEBASS-DB is a collection of five high-quality listening test datasets including test signals, reference signals and listener ratings from listening tests to evaluate the BAQ of separated audio source signals. Altogether, over 11,000 ratings of over 900 rated audio signals were collected as part of three research projects [6–8] and made freely available to the community for research purposes. The SEBASS-DB consists of the following datasets:

- *SASSEC dataset*: Ratings from signals submitted to the Stereo Audio Source Separation Evaluation Campaign 2007 [3]

- *SiSEC08 dataset*: Ratings from signals submitted to the Signal Separation Evaluation Campaign 2008 [4]

- *SAOC DB*: Listening test results from a study on the influence of the rendering architecture on the subjective performance of source separation algorithms [7]

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 25,2024 at 14:00:18 UTC from IEEE Xplore. Restrictions apply.

- *PEASS-BAQ dataset*: Re-ratings of the PEASS [11] dataset regarding Basic Audio Quality
- *SiSEC18 dataset*: Ratings from separated source signals from the SiSEC18 MUSDB18 dataset [10]

Each dataset is described in the following in a separate paragraph. Statistical key figures are presented together in Table 1. These key figures are given for each dataset:

- *Number of systems*: The number of separation systems
- *Target source signal*: The type of the target source signal to be extracted : speech or other (instruments, vocals)
- *Signal mixture*: The signal mixture the target sources are extracted from (mono or stereo)
- *Overall rated signals*: The number of rated signals in the listening test
- *Total ratings*: The overall number of performed ratings in the listening test
- *Ratings per signal*: The minimum, average and maximum number of ratings per signal in the listening test
- *Signal duration*: The duration of the signals under test in seconds
- *Listeners*: The listening expertise of the listening test participants (expert or naive)
- *Average confidence interval*: The average of the confidence intervals calculated from the ratings of the listeners for each rated item
- *Standard Deviation confidence interval*: The Standard Deviation thereof
- *Average interquartile range*: The average of the 25% and 75% interquartile range calculated for each rated item. 25% and 75% confidence intervals are pooled for the mean calculation and not calculated separately
- *Standard Deviation interquartile range*: The Standard Deviation thereof

3.1. SASSEC dataset

The SASSEC dataset contains listener ratings from signals submitted to the instantaneous mixtures separation task from the Stereo Audio Source Separation Evaluation Campaign 2007 (SASSEC) [3]. Additional source signal estimates were generated using MPEG Spatial Audio Object Coding (SAOC) [13, 14] and also rated. SAOC served as an example of an informed source separation based system in comparison to the other, blindly separated signal estimates in the test.

The listening test was split into two sessions to avoid fatigue of the listeners. All participants were expert listeners and had previously taken part in several listening tests rating the Basic Audio Quality of coded audio material. All items under test were scaled before presenting them to the listener, as proposed by Lee [15], so that the level of the target source signal estimate in the separated signal is approximately equal to that of the undistorted reference target source signal.

3.2. SiSEC08 dataset

The SiSEC08 dataset contains ratings of signals submitted to the Signal Separation Evaluation Campaign 2008 [4]. The dataset differs from the SASSEC dataset mainly in the sense that the signals the listeners rated in the listening test were separated by other source separation methods. Thus, source signal estimates from the same audio mixtures, but separated with different algorithms than in the SASSEC dataset were rated by the listener. As with the SASSEC dataset, the listening test was divided into two sessions, only experienced listeners participated, and all items were scaled prior to presentation as suggested by Lee [15].

3.3. SAOC DB

The SAOC DB contains the results of a listening test to evaluate the influence of the time/frequency (t/f) rendering architecture on the perceived quality of the BASS algorithms [7]. 14 separated source signal estimates from 9 algorithms, reflecting different types of source separation algorithms and different types of t/f processing architectures, were selected and presented to the listener. These signals were also rated in the SASSEC listening test. In order to assess the influence of the applied t/f rendering architecture on the perceived quality of the separated source signals, the listening test included additional material which was generated by using these signal estimates to drive an enhanced time/frequency rendering architecture. The rendering architecture offered by MPEG Spatial Audio Object Coding (SAOC) [13, 14] was selected, as it was designed to avoid artifacts caused by signal manipulations in the t/f domain. The material was generated by applying SAOC in a post-processing step on the separated source signals. In this way, the same basic separation core algorithm was applied together with different t/f rendering architectures and rated in this listening test.

The listening test was divided into 3 sessions, each session consisting of 14 trials. In each trial, the participants rated separated signals from 3 out of the 9 algorithms with and without post-processing, the source si1gnal (hidden reference) and the signal mixture (anchor signal) in comparison to the known original source signal (reference).

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 25,2024 at 14:00:18 UTC from IEEE Xplore. Restrictions apply.

	SASSEC	SiSEC08	PEASS DB	SiSEC18	SAOC
Number of systems	13 (11)	11 (9)	8 (4)	36 (34)	20 (18)
Target source signal (%): speech / other	57 / 63	57 / 63	50 / 50	0 / 100	57 / 63
Signal mixture (%): Mono / Stereo	0 / 100	0 / 100	20 / 80	0 / 100	0 / 100
Overall rated signals	182 (154)	154 (126)	80 (40)	192 (144)	336 (252)
Total ratings	2730 (2310)	2156 (1764)	560 (280)	2696 (2022)	3056 (2292)
Ratings per signal min/avg/max	15 / 15 / 15	14 / 14 / 14	7 / 7 / 7	10 / 14 / 19	6 / 9 / 12
Signal duration (s)	10	10	5	10	10
Listeners	Experts	Experts	Experts	Experts & Naives	Experts
Average confidence interval	5.6 (6.1)	4.9 (5.2)	9.4 (10.9)	6.86 (7.7)	5.8 (6.6)
Standard Deviation confidence interval	2.4 (1.8)	2.0 (1.3)	4.0 (2.0)	3.1 (2.0)	3.1 (2.4)
Average interquartile range	14.7 (16.1)	12.4 (13.0)	18.3 (21.2)	20.5 (23.6)	13.9 (16.8)
Standard Deviation interquartile range	6.8 (5.7)	6.4 (5.2)	8.3 (5.6)	11.4 (9.3)	8.9 (7.8)

Table 1. Overview SEBASS-DB. Numbers in round brackets are without hidden reference and anchor signals.

3.4. PEASS-BAQ dataset

An additional listening test was carried out on the original PEASS dataset [11] and released as the PEASS-BAQ dataset. The original listening test results were used by Emiya et al. to calibrate PEASS [11], a toolkit for estimating the perceived quality of separated source signals regarding the influence of artifact, interferer and target source related distortions. The signals were re-rated since the original listening test instructions differ slightly from those used to perform the listening tests for the SEBASS-DB datasets. The participants rated in the original listening test the "global quality" instead of Basic Audio Quality. Further, the listeners were instructed to rate the worst item over all items with zero, which is not prescribed according the MUSHRA listening test procedure.

The PEASS-BAQ set contains ratings of separated signals from instantaneous but also from anechoic, convolutive and professional mixes. Moreover, 3 different types of anchor signals are also included in the dataset and were rated in the listening test: Artifact related and target source distortion related anchors in addition to the signal mixtures.

3.5. SISEC18 dataset

In the course of the audio source separation campaign 2018 (SiSEC2018) [10], the organizers released the MUSDB18, a dataset of 150 full-length tracks divided into training (100 tracks) and test set (50 tracks). Participants of the SiSEC2018 Music Separation Task (SiSEC2018 MUS) were asked to separate each of the 50 tracks from the test set into the five components of "vocals", "accompanimet","drums ","bass" and "other", for a total of 50×5 separation tasks for each participant. The submitted signal estimates can be downloaded from the SiSEC MUS 2018 homepage together with signal estimates from oracle estimators, in total mostly 30 signal estimates per separation task. A subset from these $\approx 50 \times 5 \times 30$ signal estimates was selected and rated in a listening test resulting in the present SISEC18 dataset of listening test results. In selecting the items, care was taken to ensure that the resulting subset contained estimates from various types of source signals of different quality and from different separation systems. Additionally, a subset of about 40 signals from the SiSEC08 dataset was selected and also rated in the listening test. Experienced and naive listeners took part in the listening test.

4. CONCLUSIONS

The SEBASS-DB [1] is presented, which is a freely available meta database of five high quality MUSHRA-based listening test results on the Basic Audio Quality of (blindly) separated audio source signals. It contains altogether over 11,000 ratings, mostly from expert listeners, from over 900 rated signals and is currently the largest one in the field of Blind Audio Source Separation known to the authors. Both listening test results and the according test and reference signals can be downloaded for each of the listening tests to support the community in the development and improvement of perception based computational measures in the field of audio source separation. We made the SEBASS DB freely available for research purposes with the hope and intention of furthering the sharing of such data begun by Emiya et al. with PEASS [11].

5. REFERENCES

[1] Thorsten Kastner and Jürgen Herre, "SEBASS-DB," https://www.audiolabs-erlangen.de/resources/2019-

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 25,2024 at 14:00:18 UTC from IEEE Xplore. Restrictions apply.

WASPAA-SEBASS/, Oct. 2019.

[2] ITU-R BS.1534-3, “Recommendation ITU-R BS.1534-3 Method for the subjective assessment of intermediate quality level of audio systems,” 2015.

[3] Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian P. Rosca, “First stereo audio source separation evaluation campaign: Data, algorithms and results,” in *7th International Conference on Independent Component Analysis and Signal Separation (ICA07)*, London, United Kingdom, Sep 2007, pp. 552–559.

[4] Emmanuel Vincent, Shoko Araki, and Pau Bofill, “The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation,” in *8th International Conference on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brazil, Mar 2009, pp. 734–741.

[5] F.-R. Stoeter, A. Liutkus, and N. Ito, “The 2018 signal separation evaluation campaign,” in *14th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA 2018)*, Surrey, United Kingdom, Jul 2018.

[6] Thorsten Kastner, “Evaluating physical measures for predicting the perceived quality of blindly separated audio source signals,” in *Audio Engineering Society Convention 127*, New York, Oct 2009, number 7824.

[7] Thorsten Kastner, “The influence of the rendering architecture on the subjective performance of blind source separation algorithms,” in *Audio Engineering Society Convention 127*, New York, Oct 2009, number 7898.

[8] Thorsten Kastner and Jürgen Herre, “An efficient model for estimating subjective quality of separated audio source signals,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’19)*, New Paltz, New York, USA, October 2019.

[9] Emmanuel Vincent, Maria G. Jafari, and Mark D. Plumbley, “Preliminary guidelines for subjective evaluation of audio source separation algorithms,” in *Proceedings of the ICA Research Network International Workshop*, 2006, pp. 93–96.

[10] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, “The MUSDB18 corpus for music separation,” Dec. 2017.

[11] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, “Subjective and objective quality assessment of audio source separation,” in *Transactions on Audio, Speech, and Language Processing*. IEEE, 2011, vol. 19 (7), pp. 2046 – 2057, PEASS Software Version 2.01 is used. Code and subjective database are available at http://bass-db.gforge.inria.fr/peass.

[12] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre, “webMUSHRA — a comprehensive framework for web-based listening tests,” *Journal of Open Research Software*, vol. 6, 02 2018.

[13] Jürgen Herre and Sascha Disch, “New concepts in parametric coding of spatial audio: From SAC to SAOC,” in *IEEE International Conference on Multimedia and Expo*. IEEE, July 2007, pp. 1894–1897.

[14] Jonas Engdegard, Barbara Resch, Cornelia Falch, Oliver Hellmuth, Johannes Hilpert, Andreas Hoelzer, Leonid Terentiev, Jeroen Breebaart, Jeroen Koppens, Erik Schuijers, and Werner Oomen, “Spatial audio object coding (SAOC) - the upcoming mpeg standard on parametric object based audio coding,” in *124th AES Convention Amsterdam*, May 2008.

[15] Nakjin Choi, Inseok Heo, Mingu Lee, and Koeng-Mo Sung, “On evaluation of blind audio source separation,” in *Audio Engineering Society Conference: 34th International Conference: New Trends in Audio for Mobile and Handheld Devices*, Jeju, Korea, Aug 2008, number 19.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 25,2024 at 14:00:18 UTC from IEEE Xplore. Restrictions apply.