BinauRec: A dataset to test the influence of the use of room impulse responses on binaural speech enhancement

Louis Delebecque, Romain Serizel
Université de Lorraine, CNRS, Inria, Loria
F-54000 Nancy, France
louis.delebecque@loria.fr, romain.serizel@loria.fr

***Abstract*—Thanks to spatial information contained in reverberated signals, multichannel speech enhancement (SE) algorithms are able to outperform single channel systems. Reverberated signals are often generated from simulations of room impulse responses (RIRs). However, the influence of such methods on SE quality has not been investigated so far. In this paper, we propose a dataset for binaural SE, composed of real recordings, measured and simulated RIRs for the same audio scenes. Measurements are realized with and without a dummy head. This dataset is used to estimate the relevance of evaluating SE algorithms on data generated from RIRs. Results show that the head shadow effect must be taken into account and that using simulated RIRs can lead to underestimate or overestimate SE quality depending on the spatial settings.**

***Index Terms*—multichannel speech enhancement, dataset, room impulse responses,**

I. Introduction

Compared to single channel SE systems, multichannel systems have shown to improve performances [1], [2]. Indeed, single channel systems can reduce efficiently the noise level but often introduce some distortions on speech signal. Multichannel systems are able to overcome this limitation, thanks to spatial information contained in the reverberated mixture signals.

Recently, deep neural network (DNN)-based solutions have become very popular in different audio related tasks, due to the important improvement they enable. Such solutions need to be trained on a large amount of data. In addition, the training set has to cover a large variety of cases to ensure good performances during the evaluation. For those reasons, DNN-based SE systems are often trained and evaluated on simulated signals where the reverberation is obtained from basic RIR simulators based on image source method (ISM) [3]. This approach allows for quickly generating data for a large diversity of rooms and spatial scenarios. However, it can rely on a poor physical modeling that might affect the reliability of the SE evaluation.

This work was made with the support of the French National Research Agency, in the framework of the project DiSCogs "Distant speech communication with heterogeneous unconstrained microphone arrays" (ANR-17-CE23-0026-01). Experiments presented in this paper were partially carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000).

Other approaches have been proposed improve the physical behavior of the ISM while remaining very flexible [4]–[6]. Another approach consists in measuring RIRs. This solution allows for capturing the whole complexity of room effect. However, collecting a large amount of data that includes an adequate diversity is time expensive. Moreover, the method used to measure RIRs may introduce some artifacts in the estimation. Finally, direct recordings of mixture, along with speech and noise signals, allow providing more reliable data but is even more time expensive and less flexible.

Among dataset available for multichannel SE tasks, some include recordings of real noise signals along with non-reverberant speech signals (WHAM!) [7] or with simulated speech signals (WHAMR!) [8]. The dataset of the clarity challenge [9] is simulated using binaural RIRs obtained with a hybrid simulator based on ISM/ray tracing technique [4] along with a dataset of head-related transfer functions to model the head effect.

To our knowledge, the influence of the method used to generate evaluation data has not been studied yet, and no dataset is available for this purpose. In this paper, we propose an open-source dataset that allows for studying the suitability of using synthetic reverberated signals to evaluate binaural SE algorithms. The dataset is built using the three different approaches described above for same spatial settings. It includes recordings of separated speech, noise and mixture signals along with measured and simulated RIR.

The dataset is described in section II. SE systems chosen to evaluate the approaches used to generate reverberated data, are detailed in section III. Finally, results are presented and discussed in section IV.

II. BinauRec Dataset

BinauRec[1] is a parallel dataset that represents the same audio scenes using 3 different methods: signals used for evaluation are either directly recorded, convoluted with measured RIRs or convoluted with simulated RIRs.

Signals and RIRs measurements are performed using a dummy head. The measurements are repeated without the head in order to study the head shadow effect on SE performances.

[1]BinauRec dataset is available at https://zenodo.org/record/7256984

A. Clean speech and noise signals

Clean speech samples are taken from LibriSpeech [10]. Two different types of noise are used as interference, the first one is every day life noise downloaded from Freesound [11] and the second one is speech-shaped noise. BinauRec is a test set, composed of 1800 audio clips with an average duration of 9.43 s. For each audio clip, clean speech is associated to both types of noise, which yields a total duration of 9.43 hours. For each audio scene, speech and noise are recorded separately as well as the noisy mixture.

B. Room, sources and receivers

The room has rectangular shape with the following dimensions: a length of 6.62 m, a width of 2.57 m and a height of 2.60 m. The reverberation time over 60 dB is measured to 0.20 s. The dummy head (GRAS 45BB KEMAR Head & Torso) is located at 2.27 m from the wall along the length axis and in the middle of room along the width axis. Loudspeakers (Klein & Hummel O 110) and the ears of the dummy head are set at 1.48 m from the floor. Audio signals are recorded using the portable hearing laboratory (PHL) [12] along with behind-the-ear (BTE) hearing aids shells. Each BTE, that includes two omnidirectional microphones, is positioned on each ear of the dummy head.

C. Input SIR and spatial scenarios

BinauRec is built from 6 spatial scenarios. Each scenario consists in one speech source and one noise source playing on separated loudspeakers. The loudspeakers are positioned at a distance of 1 m from the dummy head, either in front of the head or at a ± 45 or ± 90 degree, as represented in Figure 1. A relative calibration is performed using a pseudo-random noise signal with a Gaussian distribution, to ensure that all the loudspeakers deliver the same acoustic level above the dummy head. At this specific point, the level of speech signal is set to 70 dB SPL, while the level of the noise signal is adjusted to obtain the following signal to noise ratios (SNRs): -5, 0 and 5 dB. A set of recordings is realized without the dummy head, with BTE remaining at the exact same position as when the head was present, for a SNR of 0 dB.

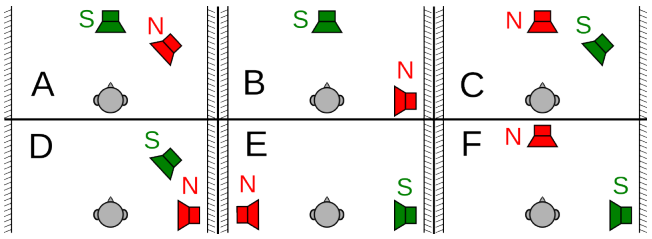

Fig. 1. Spatial scenarios of BinauRec dataset.

D. RIR measure and simulation

RIRs from loudspeakers to BTE microphones are estimated using synchronized swept-sine method [13], in both condition with and without the dummy head. Simulated RIRs are obtained from Python toolbox Pyroomacoustics [14], using the parameters described in section II-B. The option based on hybrid ISM/ray tracing is not used, except in subsection IV-B.

III. Experimental setup

A. Algorithms

Tango is a distributed multichannel SE solution designed for spatially unconstrained microphone arrays [15]. Here, each BTE is considered as a node with 2 microphones. Tango is based on 2 filtering steps: in the first step, only mixture signals from the local device are processed while the second step exploits filtered signals from the other devices as well. In each step a time-frequency masks associated to the speech signal is estimated by a DNN, the resulting estimation is then used to compute a speech distortion weighted multichannel Wiener filter (SDW-MWF), that is finally applied to input mixtures. In the current study, we use a CNN and the filter is based on generalized eigenvalue decomposition (GEVD) [16], similarly as in Delebecque et al. [17].

In addition to Tango, we evaluate the different datasets on FaSNet [18], that is another state-of-art solution for spatially unconstrained microphone arrays, to investigate the relevance of using recorded signals or RIRs for SE algorithm evaluation. FaSNet is a two stage DNN based multichannel filtering approach. In the first stage all the channels are compared to a reference channel while in the second stage the beamformer coefficients are estimated and applied to all the channels.

Tango and FaSNet DNNs are both trained using the synthetic dataset described in Furnon et al. [15]. RIRs are simulated using Pyroomacoustics with the ray tracing option disabled. The dimensions of rooms are randomly selected between 3 and 8 m for the length, between 3 and 5 m for the width and between 2.5 and 3 m for the height. In the same way, the reverberation times of rooms are chosen between 0.15 and 0.4 s. One speech source, one noise source and 16 microphones divided into 4 nodes are placed randomly in the room, with the only constraint that they all should be distant of at least 50 cm from each other and from the walls. Tango outputs one signal per hearing side whereas for FaSNet algorithm, binaural signals are obtained by applying the algorithm twice, first using the front left channel as the algorithm reference and then with front right channel.

B. Metrics

The performance of these algorithms is estimated using source to interferences ratio (SIR) and source to artifacts ratio (SAR) metrics [19], to account for noise reduction and artifact introduction, along with the input SIR to represent the balance between speech and noise in the input mixture. The reference signals used to compute the metrics are the reverberated speech and noise signals, and we use the implementation provided by Scheibler [20].

IV. Results and discussion

In order to determinate the suitability of using synthetic data to evaluate SE algorithms, we investigate for the differences in the performances obtained using the different approaches.

As our RIR simulator do not account for the head, we expect to observe some variations between measured and simulated cases. The Figure 2 compares the performances of both algorithms obtained on recorded signals, on signals convoluted with measured RIRs and with simulated RIRs.

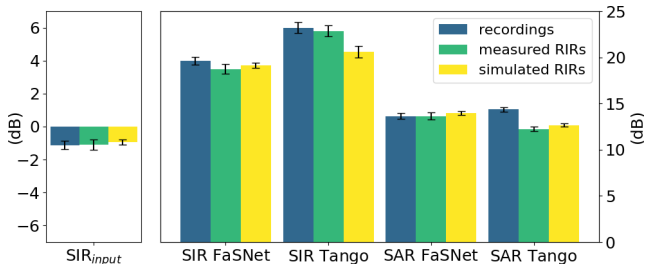

Fig. 2. Mean SE performances obtained for both ears and for all scenarios, on recorded signals, on signals convoluted with measured RIRs and on signals convoluted with simulated RIRs. Signals and RIR measurements are realized with the dummy head.

For FaSNet, comparison between those 3 approaches shows no significant differences. In the case of Tango, we observe some small variations (a lower SIR value on simulated RIRs and a higher SAR value on recorded signals). This might lead to conclude that using measured or simulated RIRs do not change the performances compared to what we obtain using recorded signals. However, those results do not allow to highlight the variations across each side of the head or across each spatial scenarios.

A. *Head shadow effect*

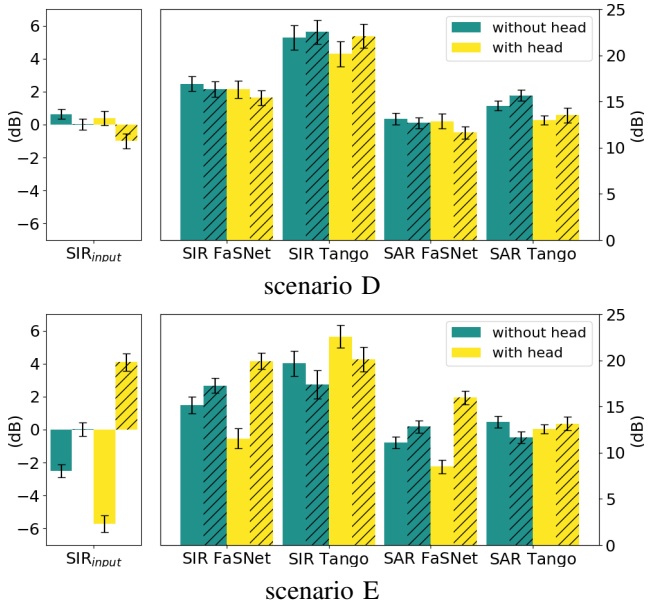

Fig. 3. Binaural metrics for FaSNet and Tango algorithms obtained on recorded signals with and without the dummy head for 2 spatial scenarios. Plain bars and hatched bars represent respectively the metric values obtained at left and right ear.

Figure 3 compares the binaural SE performances obtained on recorded signals with and without the dummy head for 2 spatial scenarios. As expected, adding the dummy head tends to decrease the input SIR at the ear that is the closest to the noise source for both scenarios. In the case of the scenario E, adding the head causes an increase in input SIR at the ear located at the opposite side as well. Those variations are due to the head shadow effect. For each hearing side, the head masks the sound source located behind. The analysis of the output metrics shows that, for FaSNet, the observed variation in the input SIR is correlated with a significant variation in both SIR and SAR metrics. This effect is not observed in the case of Tango algorithm. FaSNet is based on a reference channel that must be chosen by the user. In the case where the reference signal has a poor quality, e.g. a low input SIR, the initial noise reduction step will produce a poor quality signal and the algorithm will achieve poor SE performances. Tango performs regardless the reference channel thanks to its distributed structure and the fact that the SDW-MWF is based on GEVD. The results of this section suggest that, for some spatial scenarios, the head shadow effect causes significant variations in binaural SE performances and hence, should be taken into account when evaluating SE in binaural setups.

B. *Influence of synthetic data*

The current section aims to study the influence of using synthetic data instead of real data to evaluate SE algorithms. Since the simulated RIRs do not account for the dummy head, we compare the performances obtained using simulated RIRs, generated with and without the ray tracing option enabled, with the ones obtained using recorded signals and measured RIRs without the dummy head. The figure 4 presents the results.

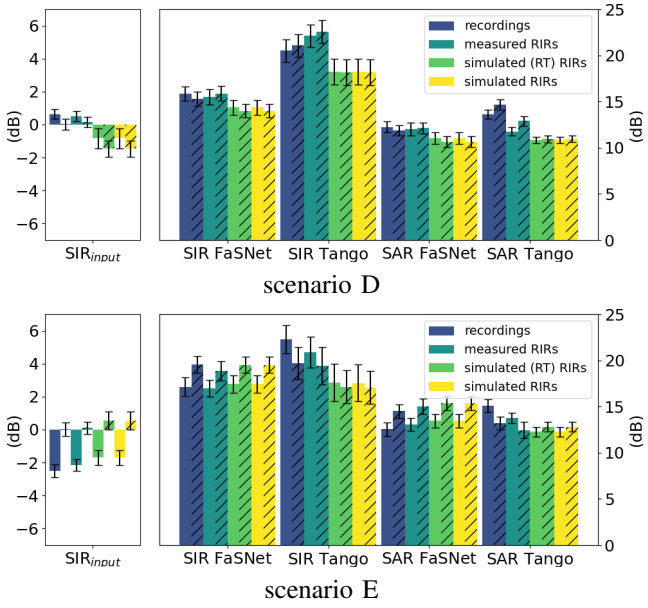

Fig. 4. Binaural metrics for FaSNet and Tango algorithms obtained without the dummy head for 2 spatial scenarios. Plain bars and hatched bars represent respectively the metric values obtained at left and right ear. RT denotes the usage of ray tracing option of the RIRs simulator.

The first interesting point to notice is that, using simulated RIRs, the input SIR may vary significantly from the values obtained from recorded signals and measured RIRs. This is the case, for example, for the scenario D but the same effect is observed for scenarios A, C and F as well. For all scenarios, values of input SIR obtained with measured RIRs are closer to the one obtained on recorded signals, compared to the ones from the simulated RIR. As observed in section IV-A, variations of input SIR are correlated with the final SE performances for SIR and SAR metrics. In the case of scenario D, using simulated RIRs degrades the performances for both algorithms. Note that in the case of scenario A, simulation gives a higher input SIR along with an improvement in SE performances. For all spatial scenarios, input SIR values and SE performances obtained with the ray tracing option enabled, do not differ from the ones obtained using the classic RIRs simulator based on ISM. Therefore, results obtained from ray tracing simulations will not be presented in the remainder of the paper.

The current experiment shows that, using simulated RIRs may lead to underestimate or overestimate binaural metrics of a given SE algorithm, depending on the spatial settings. If that is possible, using measured RIRs gives, in most cases, performances that are closer to the one obtained on real data.

C. Correction of the input SIR

In this section, we want to determine if the variations in SE performances observed during previous experiments are only due to input SIR variations or if there is some other effects that are independent to input SIR. To achieve this goal, we propose to correct the input SIR of signals obtained from measured and simulated RIR, before to applying SE algorithms and comparing performances. This correction is made by adjusting the gain of reverberated speech and noise signals to get, for each audio scene, the same input SIR value as the recorded one. It is important to note that, for a given spatial scenario, the correction gain varies for each signal, which means that the correction is signal dependent and can not be achieved by a simple calibration using one gain per loudspeaker. Figure 5 presents the results.

For FaSNet, once the input SIR values are close for all approaches, the performances obtained from measured and simulated RIRs do not differ from the ones obtained on recorded signals. This means that the variations observed without correction for the SAR (see scenario D in Figure 4), are explained by input SIR variations.

In the case of Tango algorithm, Figure 5 shows that using signals from simulated RIRs instead of real signals degrade SE quality by introducing some distortions which is indicated by lower SAR values. This effect is significant for scenarios B, D and E where the noise source is located close to the walls. In these cases, Tango algorithm exploits some information that are contained in recorded signals but missing in signals obtained from simulated RIRs. The origin of this effect could lie in disparities between the frequency content of recorded and simulated signals, that might be caused by

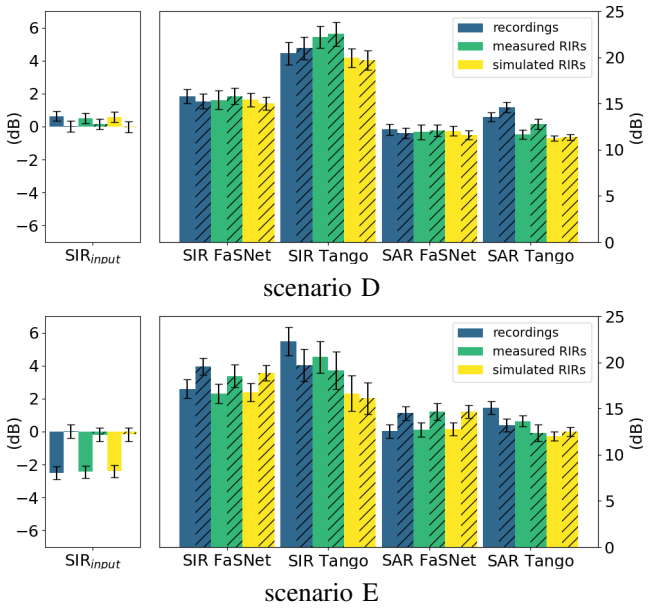

Fig. 5. Binaural metrics for FaSNet and Tango algorithms obtained without the dummy head and with input SIR correction for 2 spatial scenarios. Plain bars and hatched bars represent respectively the metric values obtained at left and right ear.

steps that are poorly or not modeled in the simulation, like the directivity patterns of loudspeakers or microphones, wave reflection against the walls or frequency bandwidth of the microphones.

Those results show, that the variations in input SIR between real and simulated data are responsible for variations in final SE performances. However, for some specific spatial settings, algorithms may exploit information that are missing in simulated data. Once again, evaluating SE algorithms on simulated data only might not lead to reliable conclusions regarding the algorithms performance in real scenarios.

V. Conclusion

In this paper, we propose an open-source dataset for the evaluation of binaural SE algorithms. It includes reverberated speech, noise and mixture signals, RIRs measured with and without a dummy head as well as RIRs simulated using ISM. The dataset is used to study the suitability of using synthetic data for evaluating binaural SE algorithms. The comparison between SE performances obtained on recorded and synthetic data, for 2 different algorithms shows that basic RIR simulations based on ISM are not reliable enough. Depending on the spatial settings, SE performances can be underestimated or overestimated.

Additionally, experiments demonstrate that the head shadow effect can affect dramatically the binaural SE performances and should be taken into account in the method used to generate the evaluation data. Finally, using measured RIRs allows to obtain more reliable data and should be considered as an interesting flexible alternative to generate evaluation set.

Future research directions could include checking if using a more advanced RIR simulation tool that takes into account for the head [9], loudspeaker directivity or that is based on a more realistic acoustic model for reproducing accurately SE performances obtained on real data.

VI. Acknowledgment

The authors would like to thank Chaslav Pavlovic for providing us with the portable hearing laboratory platform (PHL)[2] that was used to record the dataset samples, as well as Balbine Maillou and Joël Ducourneau, from University of Lorraine and LEMTA for accessing the dummy head platform and for helpful support in its use.

References

[1] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE/ACM TASLP*, vol. 19, no. 7, pp. 2159–2169, 2011.

[2] E. Ceolini, I. Kiselev, and S.-C. Liu, "Evaluating multi-channel multi-device speech separation algorithms in the wild: a hardware-software solution," *IEEE/ACM TASLP*, vol. 28, pp. 1428–1439, 2020.

[3] J. B. Allen and D. 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.

[4] D. Schröder and M. Vorländer, "Raven: A real-time framework for the auralization of interactive virtual environments," in *Forum acusticum*. Aalborg Denmark, 2011, pp. 1541–1546.

[5] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[6] D. Poirier-Quinot, M. Noisternig, and B. Katz, "Evertims: Open source framework for real-time auralization in vr," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, 2017, pp. 1–5.

[7] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[8] M. Maciejewski, G. Wichern, and J. Le Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[9] S. Graetzer, J. Barker, T. Cox, M. Akeroyd, J. Culling, G. Naylor, E. Porter, R. Viveros Munoz *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2. International Speech Communication Association (ISCA), 2021, pp. 686–690.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," *IEEE ICASSP*, pp. 5206–5210, 2015.

[11] F. Font, G. Roma, and X. Serra, "Freesound technical demo," *ACM International Conference on Multimedia (MM'13)*, pp. 411–412, 2013.

[12] C. Pavlovic, R. Kassayan, S. Prakash, H. Kayser, V. Hohmann, and A. Atamaniuk, "A high-fidelity multi-channel portable platform for development of novel algorithms for assistive listening wearables," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2878–2878, 2019.

[13] A. Novak, P. Lotton, and L. Simon, "synchronized swept-sine: theory, application, and implementation," *journal of the audio engineering society*, vol. 63, no. 10, pp. 786–798, october 2015.

[14] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *IEEE ICASSP*, Apr 2018.

[15] N. Furnon, R. Serizel, S. Essid, and I. Illina, "Dnn-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays," *IEEE/ACM TASLP*, vol. 29, pp. 2310–2323, 2021.

[16] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants," *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 785–799, 2014.

[17] L. Delebecque, R. Serizel, and N. Furnon, "Towards an efficient computation of masks for multichannel speech enhancement," Mar. 2022, working paper or preprint. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03604983

[18] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 ASRU*. IEEE, 2019, pp. 260–267.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] R. Scheibler, "Sdr—medium rare with fast computations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 701–705.

[2]https://batandcat.com/portable-hearing-laboratory-phl.html