Building and Evaluation of a Real Room Impulse Response Dataset

Igor Szöke, *Member, IEEE*, Miroslav Skácel, Ladislav Mošner, *Student Member, IEEE*, Jakub Paliesek, and Jan (Honza) Černocký, *Senior Member, IEEE*

***Abstract*—This paper presents BUT ReverbDB—a dataset of real room impulse responses (RIR), background noises, and retransmitted speech data. The retransmitted data include LibriSpeech test-clean, 2000 HUB5 English evaluation, and part of 2010 NIST Speaker Recognition Evaluation datasets. We provide a detailed description of RIR collection (hardware, software, post-processing) that can serve as a "cook-book" for similar efforts. We also validate BUT ReverbDB in two sets of automatic speech recognition (ASR) experiments and draw conclusions for augmenting ASR training data with real and artificially generated RIRs. We show that a limited number of real RIRs, carefully selected to match the target environment, provide results comparable to a large number of artificially generated RIRs, and that both sets can be combined to achieve the best ASR results. The dataset is distributed for free under a non-restrictive license and it currently contains data from eight rooms, which is growing. The distribution package also contains a Kaldi-based recipe for augmenting publicly available AMI close-talk meeting data and test the results on an AMI single distant microphone set, allowing it to reproduce our experiments.**

***Index Terms*—Far-field, automatic speech recognition, room impulse response, reverberation, SineSweep, Maximum Length Sequence, noise, deep neural network, Kaldi, AMI.**

I. Introduction

AUTOMATIC speech recognition (ASR) has made tremendous improvements in the last decade and services and applications making use of close-talk speech (such as SMS dictation, personal assistants, or contact-center speech data analytics) are on the market and serving millions of customers. On the other hand, ASR from far-field microphones is far less advanced and significant research efforts are devoted to improving its performance and robustness.

Despite all the research efforts, the best one can do to obtain a decent ASR performance is to collect transcribed data from the target domain. For far-field ASR, however, this is unfeasible due to the infinity of different room configurations, microphone placements, microphone types, noise conditions, etc. *Data augmentation* — reverberation of source data using estimated or artificially generated room impulse responses (RIR) and adding real noises to simulate the environment — is therefore the most common technique to build a robust ASR [1] nowadays.

Collecting noises is easy as there are lots of public sources and the noises can also be extracted from existing speech data. On the other hand, gathering real RIRs is technically difficult and time demanding. To overcome this problem, artificial RIRs are usually used as they can be generated automatically and in large quantities. They are good enough in scenarios, where the loudspeaker and microphone are facing each other [2] (see Section V-C), but simulating RIRs for microphones, partly or fully hidden, is not widely supported by existing tools. Here, the estimation of real impulse responses is the only way.

There is also a lack of "parallel audio corpora" where both clean close-talk (ideally anechoic) speech is available together with reverberated and noised version in various environments. This parallel corpus may also be useful in scenarios such as audio enhancement, denoising, dereverberation or beam-forming.

A. Motivation and Goals

The motivation of this paper is to: a) introduce the Brno University of Technology Speech@FIT Reverberation Database (**BUT ReverbDB**) and describe the methodology of its collection; and b) compare the impact of data augmentation using either artificial or real impulse responses in scenarios with no target training data available for the development of an ASR system. BUT ReverbDB contains also data for developing and testing of far-field Speaker REcognition (SRE) system [3], [4], but this paper concentrates solely on ASR.

The BUT ReverbDB was built in order to collect a large number of various RIRs, room environmental noises (or "silences"), retransmitted speech (for ASR and SRE testing), and meta-data (positions of microphones, loudspeakers etc.). The goal is to provide the speech community with a dataset for data augmentation and distant microphone or microphone array experiments in ASR and SRE. The database is distributed under Creative Common 4.0 Attribution license (CC-BY 4.0 – free for commercial, academic, and government use) and is available on the BUT web pages.[1]

Manuscript received November 16, 2018; revised March 1, 2019; accepted May 8, 2019. Date of publication May 17, 2019; date of current version July 25, 2019. The work was supported in part by the Czech Ministry of Interior project VI20152020025 "DRAPAK", in part by the Google Faculty Research Award program, in part by the Czech Science Foundation under project GJ17-23870Y, and in part by the Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations Excellence in Science – LQ1602". The guest editor coordinating the review of this paper and approving it for publication was Dr. Shoko Araki. *(Corresponding author: Igor Szöke.)*

The authors are with the Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic (e-mail: szoke@fit.vutbr.cz; iskacel@fit.vutbr.cz; imosner@fit.vutbr.cz; xpalie00@stud.fit.vutbr.cz; cernocky@fit.vutbr.cz).

Digital Object Identifier 10.1109/JSTSP.2019.2917582

[1] https://speech.fit.vutbr.cz/software/but-speech-fit-reverb-database

So far, BUT ReverbDB contains data from 8 rooms (large, middle and small size). We placed 31 microphones in each room. The loudspeaker was usually placed at 5 different positions per room. We measured room impulse responses, environmental noise (silence) and we retransmitted **Libri-Speech Test-clean** dataset [5], **2000 HUB5 English evaluation** set[2] and also part of **NIST Speaker Recognition Evaluation 2010** dataset [6] (the availability of the HUB5 and SRE data is limited to sites that have a valid LDC license to the original data).

All loudspeaker and microphone positions are measured and stored in meta-files in Cartesian and polar coordinates, and in an absolute and relative (to the loudspeaker) way.

BUT is taking part in the "DRAPAK" project sponsored by the Czech Ministry of Interior concentrating on ASR and SRE in the security domain (including close-talk and distant microphones, listening devices, etc.), therefore, the motivation of BUT ReverbDB is to collect acoustic environments which are challenging and cannot be easily simulated. That is also why our microphones are partly placed in very unusual places.

A number of ASR experiments were performed with BUT ReverbDB — partly as a sanity check and partly to show the importance of real environment impulse responses and background noises for training data augmentation.

B. Organization of the Paper

The paper is organized as follows: the following Section II presents related work in a distant microphone ASR, existing RIR data-sets and their shortcomings. Section III summarizes approaches in estimating real and computing simulated RIRs. Section IV presents BUT ReverbDB with details and practical recommendation in Appendix A. Section V describes the first ASR experiments aimed to validate the data-set. Section VI presents training data augmentation work on AMI data – these experiments are fully reproducible as all RIRs, data and recipes are made available. Section VII concludes the paper and outlines future work.

II. Related Work

A. Automatic Speech Recognition on Reverberated Data

ASR performance heavily degrades when facing a mismatch between training and evaluation data conditions [7]. Such a mismatch can include the environment (background noise, recording conditions (microphones and rooms)) and speaker characteristics (calm speech versus shouting with Lombard effect). An obvious solution is to collect and transcribe data from the target domain. However, when ASR is used in the field, the time, effort, and cost of transcribing data for the new conditions becomes limited (as in IARPA's BABEL[3] and DARPA's LORELEI[4] projects) or prohibitive (ASpIRE challenge [8]).

Changes in room acoustics can be a significant source of mismatch (and hence an ASR word accuracy drop) as was shown in the International Computer Science Institute (ICSI) meeting room dataset [9], [10], Augmented Multi-party Interaction (AMI) meeting room corpus [11], and the Multi-channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV) corpus [12].

The ASpIRE challenge [8] addressed far-field microphone recordings of conversational speech with a very large vocabulary. The test data differed substantially from the training and development data. The ASpIRE challenge demonstrated that working continually on the same test data and making progress on that data may not guarantee robustness to data collected in new (although related) recording conditions. Reverberation was clearly important in both the development and evaluation sets; however, microphone variability was greater in development set (Mixer 6 [13]) and room variability in the evaluation set (Mixer 8). This suggests that new challenges that aim to measure system robustness need to creatively collect new test data with mismatch and then limit testing on these data until after systems are developed.

An interesting analysis of ASpIRE results [2] studied the correlation of source-to-microphone distance and ASR performance, and concluded that rather than trying to extrapolate ASR performance from simple distance metrics, one needs to also take into account the orientation of both the speaker and the microphone. This means that we do need not only data with microphones facing directly the speaker, but also other, more complicated, speaker-microphone positions.

Another paper by Ko [14] based on ASpIRE and AMI data pointed out that the performance gap between using simulated and real RIRs can be eliminated when point-source noises are added. For Ko, the trained acoustic model not only performed well in the far-field scenario, but also provided better results in the close-talking one.

The problem of robustness of ASR on distant microphones was also approached by a series of CHiME challenges. CHiME-1 [15] aimed at small vocabulary ASR (command and control) in a real living room using binaural microphones. Target speech commands were mixed into the environment noises at a fixed position using genuine room impulse responses. CHiME-2 [16] used the CHiME-1 dataset and aimed at a larger vocabulary and a more realistic mixing process accounting for small head movements while speaking. CHiME-3 [17] and CHiME-4 are designed around the popular Wall Street Journal (WSJ) corpus and feature talkers speaking in challenging noisy environments recorded using a 6-channel tablet-based microphone array. They consist of two types of data 1) "Real data" – read speech data recorded in real noisy environments (on a bus, cafe, pedestrian area, and street junction) uttered by actual talkers; and 2) "Simulated data" – noisy utterances generated by artificially mixing clean speech data with noisy backgrounds. Actually, CHiME-5 [18] aims to be the first large-scale corpus of real multi-speaker conversational speech recorded via commercially available multi-microphone hardware (Kinect and binaural microphones) in multiple homes. Speech material was gathered from a 4-people dinner party scenario in 20 homes. However, no RIRs were collected in any CHiME data collections.

In REVERB challenge [19], the goal was to evaluate different approaches to ASR and speech enhancement on simulated data

[2]https://catalog.ldc.upenn.edu/LDC2002S09

[3]https://www.iarpa.gov/index.php/research-programs/babel

[4]https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

TABLE I
COMPARISON OF PUBLICLY AVAILABLE RIR DATASETS. M2L MEANS MICROPHONE TO LOUDSPEAKER DISTANCE – * DENOTES OUR GUESS FROM PHOTOS. METADATA MEANS RECORDING PROTOCOLS INCLUDING INFORMATION AS PHOTOS, PLACING COORDINATES, TYPE OF MICROPHONES, ROOM DIMENSIONS, AND EQUIPMENT – N/A DENOTES NOT AVAILABLE

Name	# RIRs	# Rooms	RT_{60}	M2L dist. [m]	Metadata	Target	IR type
ACE	700	7	$0.34 - 1.25$	$0.5 - 4.0$	very good	DRR and RT_{60} evals	ESS
AIR	214	6	$0.12 - 0.78$	$0.5 - 10.0$	good	SE, binaural	MLS
REVERB	24	3	$0.25 - 0.70$	$0.5 - 2$	N/A	SE, ASR	N/A
RWCP	$3k$	9	$0.00 - 1.30$	$2 - 4$*	N/A	SE, ASR	TSP
BUT ReverbDB	$1.3k$	8	$0.59 - 1.85$	$0.5 - 15.0$	excellent	SE, ASR	ESS (MLS)

(WSJ artificially reverberated and noised by real world RIRs and noises) and real data (WSJ utterances read by humans in real noisy and reverberant conditions). The conclusion of the REVERB challenge [20] was "Apart from the problems of ASR techniques, concerning the data preparation stage, challenges remain in simulating acoustic data that are close to actual recordings. Developing better simulation techniques remains another important research direction since simulations can be useful to evaluate techniques and generate relevant training data for acoustic model training."

The results of Ravanelli [21] show that using real RIRs to augment the training data provides a significant improvement on the ASR Word Error Rate (WER) (using a recent deep neural network system) to the data augmentation using just artificial RIR even with setting the room parameters as close as possible to the real room.

B. Available Room Impulse Responses Sets

In the past, several attempts of collection of RIRs and environmental noises were done, either for research purposes in the field of speech enhancement, speech recognition, beamforming, acoustic environment characterization, or for smart-homes. We identified two main categories of datasets:

1) Designed for Speech Enhancement (SE) and Automatic Speech Recognition (ASR) (see Table I for details):
   - Aachen Impulse Response (**AIR**)[5] [22] database (6 types of room, with several configurations of microphone/source placing, including the binaural microphone) aims at evaluation of speech enhancement algorithms dealing with room reverberation.
   - **ACE** Corpus[6] [23] (50 microphones in 6 devices, placed in 2 setups in 7 rooms) was used in ACE challenge [24] of T_{60} and Direct-to-Reverberant Ratio estimation methods using real noisy reverberant speech. The recording devices are a mobile phone, notebook and 32-channel spherical microphone array.
   - **REVERB** challenge[7] [19] dataset (2 times 3 types of room, near and distant microphone placement, 2 microphone angles) is a common evaluation framework including datasets, tasks, and evaluation metrics for both speech enhancement and ASR. It is carefully designed to assess robustness against reverberation. It contains WSJCAM0 [12] utterances, either spoken by humans in reverberant conditions or artificially retransmitted by a loudspeaker.
   - **RWCP** Sound Scene Database[8] [25] (circular and linear microphone array placed in 9 rooms with several positions of the loudspeaker) is a data collection project that serves sound source localization, retrieval, recognition and speech recognition in real acoustical environments. It includes retransmitted phonetically balanced sentences with precise position tracking of moving loudspeaker.
2) Designed for smart-home appliances:
   - **DIRHA** project[9] [26] dataset is composed of real phonetically-rich sentences recorded in a domestic environment equipped with a large number of microphones and microphone arrays distributed in space. It has very precious material for studies on multi-microphone speech processing and distant-speech recognition. No RIRs are public.
   - **VoiceHome** [27] corpus aims at command and control, and dialog scenarios (smart-home). Reverberated and noisy speech spoken by 12 native French talkers in 4 houses (3 rooms per house) is recorded by an 8-microphone device at various angles and distances and in various noise conditions. 188 RIRs were collected, however, none are publicly available.
   - **Sweet-Home** [28] corpus also targets command and control scenario (smart home). It consists of 26 hours of speech data (French) recorded in 4 rooms (1 flat), 7 channels (2+2+2+1). No RIRs were recorded.

The *SE/ASR datasets* are mainly focused on RIR estimation and ambient noise collection. They expect the microphones to be integrated in devices placed on furniture and at a reasonable distance from the loudspeaker (up to 4 meters). The *smart-home datasets* are focused more on the command and control scenario in reverberant and noisy environments. They contain recorded sets of proprietary utterances spoken by several humans (VoiceHome and Sweet-Home in French). Microphones are expected to be integrated in walls/ceiling, as small microphone arrays. A drawback is the lack of RIRs unavailability, although they were estimated.

Overall, none of these datasets include retransmitted publicly available speech data. Next, a majority of the datasets contain several microphone arrays which limits the variability in

[5]https://www.iks.rwth-aachen.de/fileadmin/user_upload/downloads/forschung/tools-downloads/air_database_release_1_4.zip

[6]https://acecorpus.ee.ic.ac.uk/

[7]https://catalog.ldc.upenn.edu/LDC95S24

[8] http://research.nii.ac.jp/src/en/RWCP-SSD.html

[9]http://dirha.fbk.eu/English-PHdev

microphone positions. An interesting point was raised by Ravanelli [29] who found that for the ASR adaptation, variability across rooms is more important than within the room. So from our opinion it does not make much sense to place large microphone arrays in few rooms. Lastly, all datasets expect "cooperating user" by placing microphones on a furniture (smart assistants, handheld devices, etc) at a reasonable distance or integrated in walls / ceiling using small arrays with direct human to microphone visibility. In our opinion, there is a clear lack of:

1) ad-hoc microphone placement in "non-cooperative" positions (large obstacles, partly or fully hidden microphone) where the user is not even aware of the presence of a microphone.
2) retransmissions of publicly available data. Here we note that according to our experiments, this type of data can be artificially generated by reverberating the source data and adding particular noise (see Section V).
3) good metadata as many datasets contain RIRs without precise microphone / loudspeaker placing and orientation coordinates and other description. This is fine for "put all data on one heap and train a DNN" scenario, but it is not sufficient for any deeper analysis. Precise metadata may also be used for experiments comparing real and artificial RIRs (see Section III).
4) variety in acoustic environments. All available public datasets together contain RIRs from only 25 rooms (mainly offices, meeting and lecture rooms).

A good RIR dataset for SE/ASR should have a good variety over environments (rooms), microphone placing (visible, hidden), microphone types (high-end, MEMs, low-end, handheld device, integrated etc.), and precise metadata. We do not consider microphone arrays as important since there can be many variations. However, the single distant microphone still has a significant application coverage.

In conclusion, a large data set of RIRs with consistent recording protocols covering standard acoustic environments like offices, houses, corridors, cars etc., is missing. The closest RIR datasets are the ACE and AIR. Our goal — as our target application is speech data mining (ASR and SRE) from a variety of sources (table top microphones, IoTs, mobile devices, smart assistants, smart homes, but also listening devices, bugs and other non-standard microphones) — is to have RIRs from a variety of microphone positions.

III. Obtaining Room Impulse Responses

An RIR can be obtained in two principal ways: the first is to measure the environment and obtain the "real" RIR, the second is to generate it artificially by a simulation.

A. Real Room Impulse Responses

Several methods were developed to measure the real RIR. The Maximum Length Sequence (MLS) technique was first proposed by Schroeder [30]. Other techniques were suggested to reduce distortion artifacts of MLS such as the Inverse Repeated Sequence (IRS) [31]. Another method – Time-Stretched Pulses – was proposed by Aoshima [32]. Finally, a logarithmic Sine Sweep technique introduced by Farina [33] should overcome some limitation of the other ones.

We briefly summarize these techniques and refer the reader to Stan *et al.* [34] for extensive comparison with a supporting mathematical apparatus:

Maximum Length Sequence is based on the excitation of the acoustical space by a periodic pseudo-random signal [35]. The number of samples of one period of MLS signal is: $L = 2^m - 1$, where m is the order. The RIR is then calculated by circular cross-correlation between the measured output and the original MLS signal. The circular cross-correlation obviously causes a well known problem [34] – the time-aliasing error, which can be overcome by setting L longer than expected RIR measured (considering 48 kHz sampling frequency, the $m > 17$ to be on the safe side). The MLS method has a strong immunity to signals not correlated with the excitation signal, due to the MLS phase spectrum being irregular and a uniform density of probability. Any disturbing signals are "spread" uniformly along the deconvolved RIR. Using averaging as post-processing leads to the reduction of the distortions. This makes the MLS suitable for RIR measuring in an occupied room or exterior setting. On the other hand, a major drawback is in the appearance of "distortion peaks" [36]. The MLS method relies on the assumption of Linear, Time-Invariant (LTI) system. Any inherent non-linearities of the measurement system (especially the loudspeaker) are present in the RIR and appear as cracking sounds when convolved with an audio. They can be partly avoided by precise calibration (mainly the loudspeaker output level). MLS also expects input/output sampling clock synchronization [33].

Inverse Repeated Sequence reduces the "distortion peaks" drawback of MLS. The IRS excitation signal is a sequence of length $2L$, the first half is equal to MLS and the second half is inverse MLS [31]. The rest is common with the MLS method (circular cross-correlation, input/output sampling clock synchronization, immunity to disturbing signals).

Time-Stretched Pulses method reduces the distortion peaks by the expansion and compression of an impulsive signal [32]. It also relies on the assumption of LTI system. According to the spectral properties of stretched pulses, this method is not immune to disturbing signals (it cannot be used in occupied rooms [34]).

Exponencial Sine Sweep – (ESS) uses an exponential time-growing frequency sweep as the excitation signal. ESS does not rely on the LTI system assumption in contrast to the MLS, IRS, and TSP. It is possible to perform simultaneous deconvolution of the linear impulse response of the system and selective separation of each impulse response corresponding to the harmonic distortion using the excitation signal. The harmonic distortions appear prior to the linear impulse response [33]. The impulse response deconvolution process is implemented by the linear convolution of the measured output with the analytical inverse filter estimated from the excitation signal. The advantage upon MLS and IRS methods is that linear convolution overcomes time-aliasing problems. If the emitted ESS is shorter than the RIR to be measured, we just need sufficient silence to be added at the end of the ESS to recover the tail of RIR. The ESS method is perfect in rejecting the harmonic distortions as they appear prior

to the "linear" impulse response estimation. It has an excellent RIR signal-to-noise ratio. It also does not need an output level calibration. On the other hand it is not immune to disturbing signals and is suitable for quiet rooms [34].

From the experimental point of view, according to [37], the Exponential Sine Sweep (ESS) has shown robustness to changing loudspeaker output level while MLS and LSS (Linear Sine Sweep) tend to degrade the ASR WER in the presence of higher output volumes. ESS was also found robust (only 0.5% WER deterioration) when switching from expensive studio monitors to cheap PC loudspeakers.

In conclusion, the best method for our needs is the Exponential Sine Sweep [38] as it is not sensitive to output level calibration and we will use it in empty environments (non-occupied rooms). We accompany the ESS measurements with MLS to have the RIR in cases, where a microphone is placed close to a noise source and the SNR is low for this particular microphone. Our hardware setup (see Appendix A) also does not have clock signal synchronization between playback and recording device which limits the use of MLS. However, we were able to compensate this by re-sampling the recorded MLS signal (see the following section). We use MLS implementation by Thomas [39] and ESS implementation available as a free Matlab code.[10]

B. MLS — Compensation of Clock Asynchronicity

The playback / recording clock asynchronicity causes a time stretch of a recorded signal compared to the excitation one. It lead to distortions of measured RIR when applying the circular cross-correlation on the stretched signal. We conducted an experiment where we compensated the difference in clocks for the playback and recording device. We applied the cross-correlation function on the first and last recorded period of the MLS signal (we use 32 repetitions of the MLS sequence of order $m = 18$). The time shift was then applied in the re-sampling of the recorded MLS sequence in order to match the played one sample-to-sample (see Fig. 1 for RIR with and without the sampling frequency compensation).

Finally, we did an ASR experiment (see Section V for more details) where RIRs of two rooms were estimated for 31 microphones. The test data was then artificially reverberated and processed by the ASR, and we compared word accuracies of MLS- and ESS-processed test data. The average difference between compensated MLS and ESS is only 0.37% absolute on word accuracy. This shows that the compensated MLS method provides very similar RIRs to the ESS method.

Anyway, we decided not to use MLS in further experiments and stuck to ESS, but we continued recording both MLS and ESS signals and let the user choose BUT ReverbDB.

C. Artificial Room Impulse Responses

For the purpose of an artificial RIR generation, computer simulation must be performed. Approaches that have been developed may be roughly divided into two groups: wave-based and ray-based methods [40], [41]. The former techniques are designed to solve the wave equation, whereas the latter group makes use of geometrical acoustics where sound propagates in form of rays and wave nature is neglected. Wave-based methods provide more realistic results since they are inherently able to simulate sound propagation phenomena such as diffraction. However, this advantage comes at the cost of computational expense. The boundary element method [42] and finite element method [43], representatives of the wave-based group, discretize surface or volume to elements that interact according to the wave equation which is costly. This is a limitation because when augmenting training speech data, numerous different room conditions must be simulated.

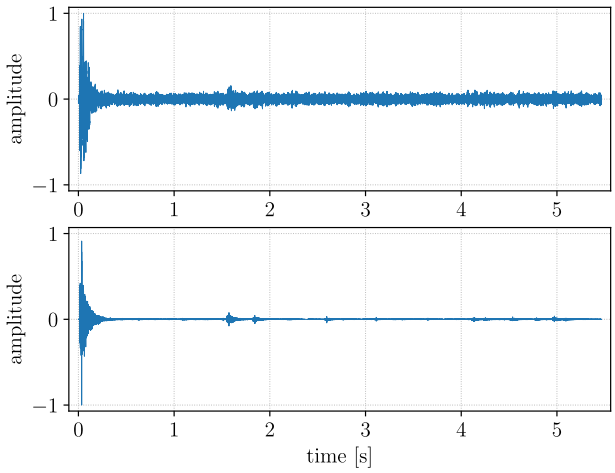


Fig. 1. Top panel shows a RIR estimated by MLS without playback and recording device clock synchronization. Notice the noise in the late reflections. Bottom panel shows RIR when the recorded MLS sequence was re-sampled to match the playback sampling frequency.

Therefore, ray-based methods are more suitable for our purpose. Image Source Method (ISM) formulated by Allen [44] and ray tracing [45] are well-known techniques based on geometrical acoustics. In ray tracing, a sound source generates multiple rays that are cast to a room at a single time instance. They propagate through free space and get reflected on walls and obstacles. Each reflection decreases ray energy according to the absorption of the material. RIR is then created using rays that passed through a receiver and their energies.

The ISM uses "unwrapping" of room geometry. Every reflection of the sound ray from a wall can be considered as a sound ray originating from a virtual source behind the wall. The sound ray energy is reduced by the wall reflection coefficient (absorption). Using this principle, the room geometry is unfolded several hundred or thousand times and appropriate virtual sound sources are placed in the space. The final RIR is a summation of delayed Dirac impulses passed through a low-pass filter (to respect the sampling theorem) and attenuated by an appropriate number of "walls" it has to reflect from.

In the speech community, the ISM is prevalent when it comes to data augmentation [46] and multiple toolkits have been created [47], [48]. To the best of our knowledge, there is no extensive study comparing ray tracing and image source

[10]http://freesourcecode.net/matlabprojects/69639/exp.sweep-and-impulse-response-in-matlab

TABLE II
LIST OF ROOMS IN THE CURRENT DISTRIBUTION OF BUT REVERBDB. THE STAR DENOTES ROOMS WITH NON-BLOCK SHAPE (FOR EXAMPLE AN "L" SHAPE). THE ROOM VOLUME IS AN APPROXIMATION. THE NUMBER OF RIRS CONSISTS OF THE NUMBER OF MICROPHONES TIMES NUMBER OF LOUDSPEAKER POSITIONS. COLUMN "RET." INDICATES NUMBER OF SPEECH DATA RETRANSMISSIONS. ROOMS USED IN THE TEST DATA EXPERIMENTS (SECTION V) ARE NOTED IN *italics*, FOUR ROOMS USED IN THE TRAINING DATA AUGMENTATION EXPERIMENTS (SECTION VI) ARE NOTED IN **BOLD**

Room ID	Dimensions [m × m × m]	Volume [m^3]	RT_{30} [s]	RIRs [#]	Ret.	Type
Q301	10.7 × 6.9 × 2.6	192	0.78	31 × 3	1	office
L207	4.6 × 6.9 × 3.1	98	0.61	31 × 6	2	office
L212	7.5 × 4.6 × 3.1	107	0.70	31 × 5	2	office
R112	4.4 × 2.8 × 2.6* 2.2 × 1.2 × 2.6*	∼40	0.59	31 × 5	0	hotel room
L227	6.2 × 2.6 × 14.2	229	1.85	31 × 11	3	stairs
CR2	28.2 × 11.1 × 3.3	1033	1.59	31 × 4	0	conf. room
E112	11.5 × 20.1 × 4.8*	∼900	1.17	31 × 2	0	lect. room
D105	17.2 × 22.8 × 6.9*	∼2000	1.13	31 × 5	1	lect. room

method for data augmentation and showing the superiority of ray-tracing. For this reason, we use the artificial RIR generator implemented by Habets [47]. It allows for setting reflection coefficients of particular walls and orientation and directional characteristics of microphones. An omnidirectional loudspeaker is considered in the simulation.

IV. BUT REVERBDB

So far, we measured 8 rooms with the majority of data processed, exported and made available. The available rooms are summarized in Table II. The volume is an approximation for non-block shape rooms. The number of RIRs is given by the number of microphones times number of loudspeaker positions. The number of retransmissions (column "Ret.") indicates how many times the speech data (LibriSpeech Test-clean, 2000 HUB5 English evaluation set, and NIST SRE 2010) was retransmitted. While RIR data was recorded for each loudspeaker position, the audio was not retransmitted for all of them, as it is a very time consuming process.

We plan to continue in the collection of BUT ReverbDB. Our goal is about 50 in-door environments including cars. We also plan to increase the number of devices by using a 2^{nd} order ambisonic microphone, MEMS microphones, tablets, mobile phones and headsets (see Appendix A for more technical details).

V. ASR TEST DATA EXPERIMENTS

In this section, we describe experiments conducted on ASR test data. Initially, this work was intended as just a set of BUT ReverbDB sanity checks, but we found that several topics are of general interest.

To begin with, we show that we are able to artificially retransmit (convolving with a RIR) test data and obtain the same word accuracy (WAC) as with the real retransmitted data. This leads to the conclusion that retransmission of acoustic data can be substituted with RIR estimation and noise recording, requiring much less time spent in the physical room.

TABLE III
DATA SOURCES USED FOR THE TEST DATA EXPERIMENTS. AUGMENTED DATA AMOUNTS ARE IN BRACKETS. WE USED A MIX OF REVERBERATION USING RIRS GENERATED BY ISM AND ADDITIVE NOISES. "TEST-SET" DENOTES DURATION (IN MINUTES) AND THE NUMBER OF SPEAKERS USED FOR ASR EXPERIMENTS IN THIS PAPER

Data	Total	Test-set	Type
SpeeCon [49]	759.4h (+996.2h)	69.9m / 15	prompted, close talk, distant mic.
Third party	641.7h (+1128.8h)	22.9m / 14	prompted, spontaneous, close talk, distant mic.
Ministry of Def.	140.0h (+247.3h)	none	spontaneous, telephone
SUM	3913.4h	92m / 39	-

We have also verified the influence of background noise on data augmentation reported in [1], [14] and confirmed that adding noise is helpful.

The influence of microphone occlusion on RIR estimation was investigated too. Theoretically, the RIR of occluded microphone can be sythesized, however, we have not yet found any tool ready to use it (see Section III). We have shown that while the ISM method is good enough for non-occluded microphone placing, when the microphone is hidden, the real RIR is a clearly superior method. This further supports the need of real RIR measuring.

We used a pre-trained Czech ASR based on stacked-bottleneck architecture [50]. The 8 kHz training data consists of 3900 hrs of telephone speech, close talk data, distant microphone data and augmented data (RIRs artificially generated by ISM and a set of publicly available noises[11]). See Table III for further details. The vocabulary and language model were derived from acoustic data transcriptions. We considered this recognizer as robust enough to provide us meaningful results. We adapted neither the acoustic model nor the language model on the test data (no speaker adaptation, no NN fine-tuning, etc.). All results are reported as word accuracy (WAC).

We selected a reasonable test-set to conduct experiments and retransmitted it in various environments. We used only clean close-talk data without reverberation and noise in the background as a source for retransmission: 92 minutes of prompted speech and phonetically balanced sentences from 39 speakers (gender and age balanced) – see Table III. We achieved 75.9% in word accuracy on the clean test-set; this is our **baseline**. We used the reference speech/non-speech segmentation in decoding the retransmitted data in further experiments, in order to suppress the influence of Voice Activity Detection (VAD) on overall results and conclusions.

We denote **Retransmit** (real retransmission) the test-set which was replayed in the particular room r and hence recorded with the room's natural reverberation and background noise by microphone c. We denote **ESS / ISM** (artificial retransmission) the test-set, where clean signal $s[t]$ was convolved with RIRs $h_{r,c}[t]$ either estimated by ESS or generated by the ISM

[11] http://freesound.org

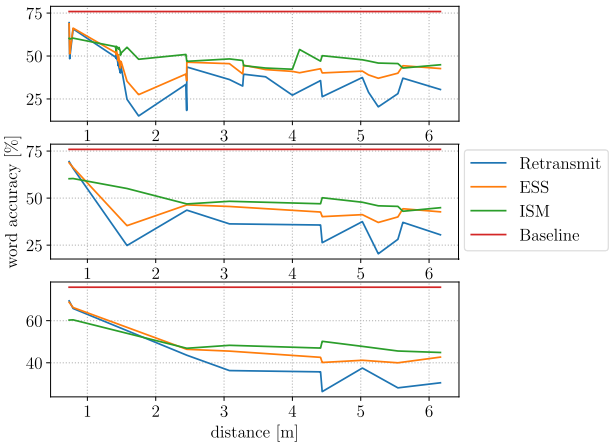


Fig. 2. Comparison of Real-Retransmitted, ESS Artificial-Retransmitted and ISM Artificial-Retransmitted test-sets in room L207. We sorted the microphones according to the distance from the loudspeaker (x-axis). The top panel shows all microphones. The middle panel shows only microphones in front of the loudspeaker (±90°). The bottom panel shows only microphones in front of the loudspeaker (±90°) with direct visibility.

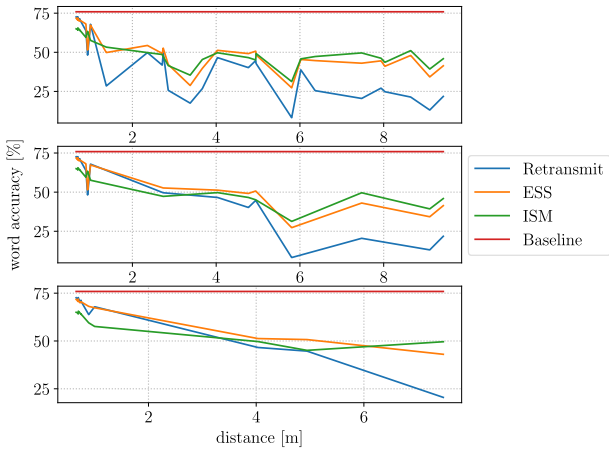


Fig. 3. Comparison of Real-Retransmitted, ESS Artificial-Retransmitted and ISM Artificial-Retransmitted test-sets in room Q301. We sorted the microphones according to the distance from the loudspeaker (x-axis). The top panel shows all microphones. The middle panel shows only microphones in front of the loudspeaker (±90°). The bottom panel shows only microphones in front of the loudspeaker (±90°) with direct visibility.

method. The resulting speech signal is then given by standard convolution:

$$s_{r,c}[t] = s[t] * h_{r,c}[t] + \alpha n_{r,c}[t + \mathit{offset}], \qquad (1)$$

In case noise $n_{r,c}[t]$ was added, the weight α is set to match the Signal-to-Noise Ratio $SNR(s_{r,c})$ estimated from the real retransmission condition in room r and microphone c using reference speech/non-speech segmentation and A-weighting function. The starting position *offset* in the noise was selected randomly, then we repeated the noise in a loop to fill the whole audio (our noise samples are 1 minute long). Data with added noise are marked with **noise** label.

In generating RIRs using ISM, we did our best to be as close as possible to the real room setup (room dimensions, loudspeaker and microphone position, microphone orientation, RT_{30} value). We estimated the RT_{30} from logarithmic decay curve [51] which was computed from an impulse response based on Schroeder integration [52]. RT_{30} was applied in ISM method using Sabin-Franklin's formula [53].

A. Simulated (ISM) vs. Real (ESS) RIRs

This section compares the influence of simulated RIR (ISM calculation) and real RIR (ESS estimation) on word accuracy. We used RIRs from two rooms and compared the Artificial-Retransmitted data to the Real-Retransmitted. As we can see from Figs. 2 and 3, there is a gap between the Real-Retransmitted and both Artificial-Retransmitted data. This is caused by missing noise (see the following section). The ESS method provides slightly more realistic RIRs to the ISM, as the word accuracies are closer to the Real-Retransmitted data.

B. Influence of Noise on Room Acoustic Simulation

We show the need of noise for test data processing in this section. We use the same data setup as in the previous section and add noise. It is a matching noise, as it comes from the particular room and microphone as mentioned earlier. As we can see from Figs. 4 and 5 compared to Figs. 2 and 3, the gap between the Real-Retransmitted and ESS Artificial-Retransmitted data almost disappears. On the other hand, there is still a gap between ISM and ESS methods showing that the artificial RIR estimation is not good enough, especially for microphones placed in non-common positions (drawer, waste bin, book shelf, etc.).

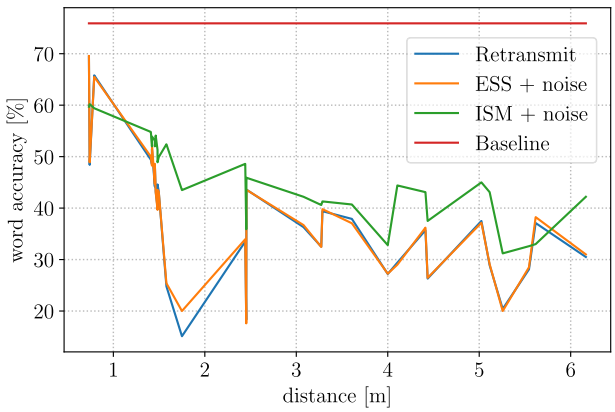


Fig. 4. Comparison of Real-Retransmitted, ESS Artificial-Retransmitted and ISM Artificial-Retransmitted test-sets in room L207. We sorted the microphones according to the distance form the loudspeaker (x-axis).

C. Microphones Occlusion

We analyzed how microphone occlusion impacts the WAC and the influence of the RIR estimation method. The impact is measured in WAC difference between Real-Retransmitted test-set and either an ESS Artificial-Retransmitted or ISM Artificial-Retransmitted test-set. We included both, only reverberation and reverberation plus additive noise. The results are shown in Table IV. It is obvious that there is no significant difference between Real-Retransmitted and the ESS Artificial-Retransmitted test-set (measured on WAC) in all microphone placement

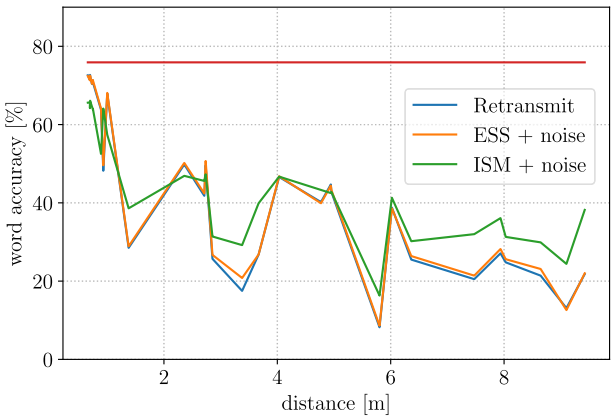


Fig. 5. Comparison of Real-Retransmitted, ESS Artificial-Retransmitted and ISM Artificial-Retransmitted test-sets in room Q301. We sorted the microphones according to the distance form the loudspeaker (x-axis).

TABLE IV
COMPARISON OF WAC DIFFERENCES BETWEEN REAL-RETRANSMITED AND ISM ARTIFICIAL-RETRANSMITTED TEST-SETS (RR MINUS ISM) AND REAL-RETRANSMITTED AND ESS ARTIFICIAL-RETRANSMITTED TEST-SETS (RR MINUS ESS) USING JUST REVERBERATION (REVERB) OR REVERBERATION AND ADDITIVE NOISE (REVERB+NOISE). THE DIFFERENCES ARE EXPRESSED AS MEAN ± STANDARD DEVIATION — A NEGATIVE NUMBER MEANS THAT ISM OR ESS PROVIDES BETTER RESULTS THAN THOSE OBTAINED WITH REAL-RETRANSMITTED DATA. THE DATA COMES FROM ROOMS L207 AND Q301 AND THE MICROPHONES HAVE FACE-TO-FACE ORIENTATION AND DIRECT VISIBILITY, ARE PARTLY BOXED (HIDDEN IN A SHELF) AND FULLY BOXED (HIDDEN IN A DRAWER). IT IS OBVIOUS THAT FOR HIDDEN MICROPHONES, ONLY ESS RIR ESTIMATION WITH ADDED NOISE PROVIDES MEANINGFUL TEST DATA

Microphone position	reverb		reverb+noise	
	$RR-ISM$	$RR-ESS$	$RR-ISM$	$RR-ESS$
Face-to-face	-1.6 ± 9.0	-4.0 ± 5.1	1.3 ± 6.4	-0.2 ± 0.5
Partly boxed	-19.8 ± 6.8	-13.0 ± 5.3	-13.0 ± 6.7	-0.2 ± 0.5
Fully boxed	-21.5 ± 7.8	-11.9 ± 5.8	-14.8 ± 8.0	-2.2 ± 2.2

conditions. Note however a stronger degradation of ISM compared to ESS Artificial-Retransmitted in occluded microphones.

VI. ASR TRAINING DATA AUGMENTATION

In a real-world ASR, one has to train ASR which is able to cope with a particular channel (far-field microphone in our case) without having target training data. As mentioned in Section II, the best performing technique is data augmentation. We used an AMI dataset [54] for this experiment; our unseen channel was the *Single Distant Microphone* – SDM and the only data available was *Individual Headset Microphone* – IHM. Our goal is to test data augmentation of AMI data using BUT ReverbDB and to investigate suitable reverberation techniques. We do not run extensive experimentation with noises; we use just the noises from BUT ReverbDB and add them to the training audio.

This set of experiments is inspired by Ko *et al.* [14]. Their work was aimed at comparing of real and simulated RIRs and adding point source noises to ASpIRE [8] and AMI datasets. On AMI, however, they only reported the impact of adding reverberated close-talk data (IHM) to the genuine distant microphone training data (SDM/MDM). We are not using SDM/MDM at all in the training.

We selected four BUT ReverbDB rooms closest to AMI meeting rooms in type and dimensions as a source of real RIRs: Q301, L207, L212, and R112 (see Table II). We did not use other public RIR sources. We generated artificial RIRs similar to the four real rooms to compare artificial versus real RIRs. Theoretically, we can generate a large number of artificial RIRs with a good chance to hit the same room configuration (dimensions, reflection coefficients, speech source and microphone position) as the target data (AMI dataset). We consider this as cheating for the time being, but we would like to perform such an experiment in our future work.

Each experiment is tagged with a used RIR set: artificial RIR (**AR**) or real RIR (**RR**) is accompanied with a number of RIRs used (**2k**, **306**, **30**). We add tag **ctXm** noting the microphone is in the range of 1 to X meters from the loudspeaker. **vis** denotes direct visibility between the microphone and the loudspeaker. Finally, **f2f** denotes "face-to-face" orientation of microphone and loudspeaker. In this way, **30.vis.ct2m.f2f* defines a set of 30 RIRs, where microphones are directly visible, closer than 2 meters and face-to-face oriented to the loudspeaker, and **306.vis.ct3m* defines a set of 306 RIRs, where microphones are directly visible and closer than 3 meters to the loudspeaker.

The training data augmentation was done in two steps: 1) reverberating the IHM audio files using selected RIRs, and 2) adding stationary noises to achieve SNR in the range of 10 to 20 dB with uniform distribution. The reverberation was done in two ways: either we convolved one whole audio file with one RIR, or changed the RIRs on-the-fly during convolution (see Section VI-D for details).

A. Baseline System Description

For acoustic models training, we used a standard AMI recipe in Kaldi [55]. The baseline system is depicted in Fig. 6 above the dashed line. First, 13-dimensional MFCC, delta and double-delta features are extracted. Cepstral mean and variance normalization (CMVN) is performed. Mono-phone GMM-HMM model is trained on a subset of the training data (about 10.8 hours of AMI IHM audio). All the data is then aligned using this system. Context-dependent tri-phone model training on the full training set (about 78 hours of audio) follows, and the data is re-aligned. Further, features are spliced together, projected to 40-dimensional space using linear discriminant analysis (LDA), and a de-correlation based on the maximum likelihood linear transform (MLLT) is applied. In the last step, the model is re-trained using speaker adaptive training (SAT). The training data is re-segmented and only the audio matching the transcriptions is selected (cleaning process) based on decoding with the GMM-HMM model and biased language model built from a reference transcript. In this way, about 7 hours of audio are discarded from the full training set. After this, the cleaned full training set is speed perturbed (original plus two speed alternations) resulting in about 210 hours of training audio. The state alignments generated by GMM-HMM system are used for DNN training. The DNNs are trained on 40-dimensional filter-bank energies along with 100-dimensional i-Vectors [56]. A time delayed

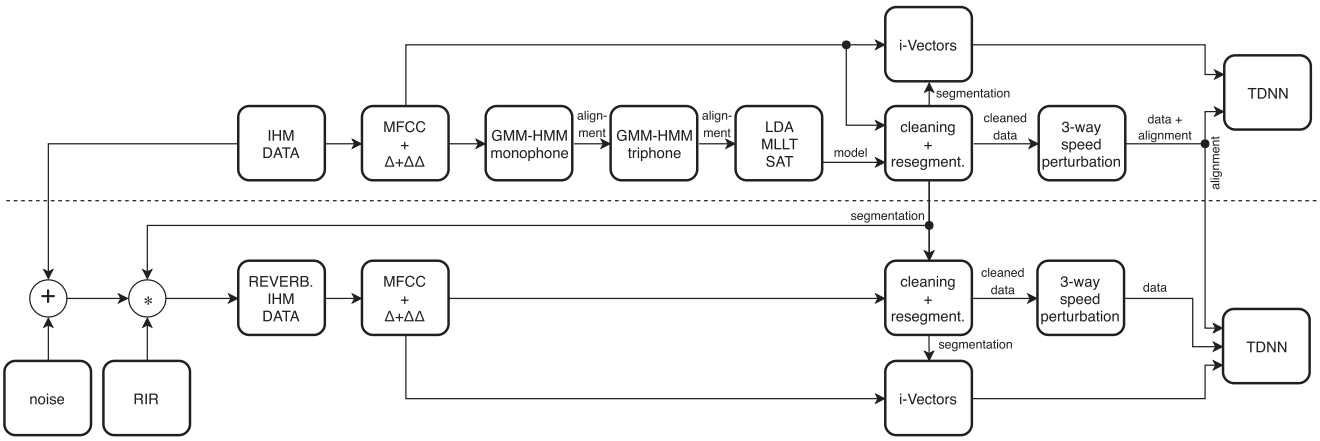


Fig. 6. Schemes of the baseline system (above the dashed line) and the modified system for reverberated data (below the dashed line).

neural network (TDNN) trained with lattice-free MMI objective is used as the final acoustic model.

B. Modifications of Kaldi Baseline

The standard AMI recipe uses the training data both for cleaning and segmentation, and for the actual acoustic model training. When using reverberated data for all these steps, we found a significant decrease in accuracy (caused obviously by worse models) and fluctuations in the amount of retained audio. Therefore, we decided to "freeze" the baseline system segmentation across all further experiments, which also implies that the same amount of training data was used (210 hours with speech perturbation). The segmentation also served for i-Vector resets (see below in Section VI-D). In the same manner, we also consistently used the baseline system alignment to train all DNN acoustic models. The modifications of a baseline system for the reverberated data are depicted in Fig. 6 below the dashed line.

C. Averaging Results

When we experimented with Kaldi AMI recipe, we found that the resulting WER in not stable enough.[12] When an experiment was run several times, we observed WER fluctuations in tenths of percent. Stability does not improve when adding more NN training iterations. As some of our experiments also differ in tenths of percents, our conclusions would not be statistically significant. That is why all results presented in this section are averages over 5 runs of ASR training (see Fig. 7 for details). We performed Student's T-test on selected pairs of systems with close average results. We concluded that 0.2% absolute difference on WER for the significance level $\alpha = 0.05$ is statistically significant (0.1% absolute difference is not significant).

D. Per Segment Reverberation

The problems of the AMI dataset are long recording and relatively small number of speakers (547). So even if we generate thousands of RIRs using ISM, only 547 are used if we apply one

[12]This is a known issue of Kaldi, probably caused by inherent nondeterminism of GPU-based matrix multiplication, as discussed by Kaldi core developers at https://github.com/kaldi-asr/kaldi/issues/2905

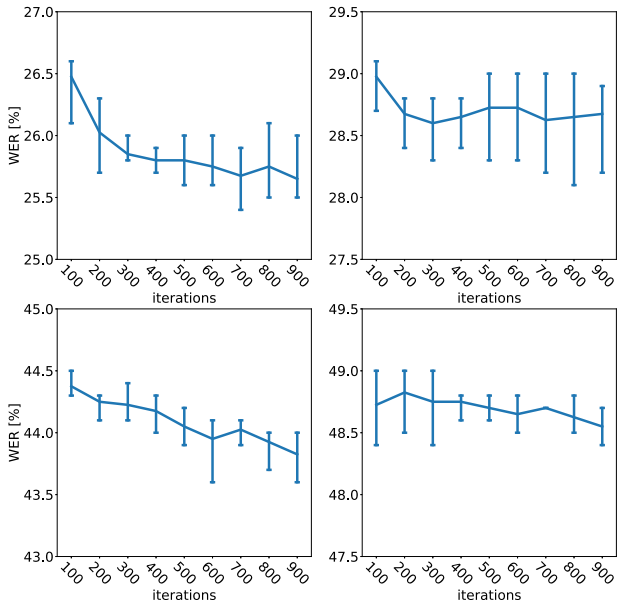


Fig. 7. Comparison of mean WER (over 5 runs). X-axis is the number of iterations in training NN, Y-axis is achieved WER for IHM (top row), SDM (bottom row), dev (left column), and eval (right column) sets. The solid left-to-right line connects means, the top and bottom lines show maximum and minimum WERs achieved for a particular run.

RIR on one whole audio file. The AMI recipe contains speaker adaptation using i-Vectors [56]. Each i-Vector is estimated on-the-fly on $2-10$ speech segments and then it is reset to ensure data variability and to prevent TDNN over-training. We modified our reverberation algorithm in order to allow changes of RIR during convolution with the audio. In the end, every speaker is reverberated with a set of RIRs and the data variety is increased compared to a single audio file reverberation.

The results (Table V) show that bringing more environmental variability per i-Vector, the reverberation decreases WER from $43.42\%/48.46\%$ to $41.80\%/47.06\%$ for SDM dev / eval set. We also conducted an experiment, where we changed the RIR only in silences longer than 3 seconds, in order to prevent artifacts in the convolution, as the RIR is 1 second long. This also "stabilizes" the channel for the i-Vector extraction and makes the i-Vectors focus on the speaker rather than acoustic environment. Here we

TABLE V
COMPARISON OF "PER SEGMENT" WITH "PER FILE" REVERBERATION. *per1seg* SETUP CHANGES RIR IN SYNCHRONY WITH KALDI I-VECTOR SPEAKER ADAPTATION. *insil* DENOTES EXPERIMENT, WHERE RIR IS CHANGED ONLY IN SILENCES LONGER THAN 3 S. COLUMN *Segm #* SHOWS NUMBERS OF SEGMENTS WITH FIXED RIR. WE RANDOMLY DRAW 547 RIRS FROM 2000 SET FOR IHM.AR547.CT3M.PERFILE SYSTEM

System	Segm #	WER [%]			
		IHM		SDM	
		dev	eval	dev	eval
ihm.AR2k.ct3m.insil	36357	21.52	23.06	41.70	46.74
ihm.AR2k.ct3m.per1seg	33312	21.44	23.02	41.80	47.06
ihm.AR547.ct3m.perfile	547	21.72	23.24	43.42	48.46

obtained another slight WER decrease from $41.80\%/47.06\%$ to $41.70\%/46.74\%$ on SDM dev / eval.

E. Room Impulse Response Passivation and Delay Compensation

Having estimated real or generated artificial RIRs, one may post-process them to achieve more consistent results and to overcome over-excitation and delays caused by the convolution. The delay in any RIR is caused by the speed of sound and can be partly compensated by measurement of microphone to loudspeaker distance. However, precise compensation is hard due to humidity and air pressure changes. Delay compensation in ISM synthesis of RIR is theoretically straightforward; the delay can be computed analytically. On the other hand, it may produce an incorrectly delayed RIR in the case of a cardioid microphone and the sound source are placed exactly behind the microphone. In this case, the direct signal is zero and we see only the reflections with larger delay than we expect from the microphone–loudspeaker distance and the speed of sound. The delay compensation is critical in data augmentation for ASR training [1], [14]. First, the labels (phonemes, senones, etc.) are aligned with the training "clean" speech data using a decoder. The clean data are swapped with the augmented (reverberated) version in the next step of training. Here, the original alignment (timing) is used with the augmented data and any time shift caused by RIR delay leads to label versus data mismatch. We denote systems with applied delay compensation by tag **shi**.

Another problem is over-excitation caused by amplifying the audio using a RIR, leading to signal clipping. To overcome this, we scale the RIR to a level which ensures that no single magnitude in the frequency response is larger than 1. The passivation has no effect when using floating-point arithmetic for convolution followed by a normalization. On the other hand, one may still face fixed-point implementations/scenarios where (latent) an overflow has a significant impact. We denote systems with applied passivation by tag **pas**.

We summarized results with RIR passivation and delay shift in Table VI. Passivation experiments are shown in the first four lines, both for artificial and real RIRs. We can clearly conclude that passivation significantly helps for artificial RIRs in the IHM condition. Passivation does not bring any significant improvement for real RIRs. This leads to the conclusion that ReverbDB RIRs are well estimated and will not cause over-excitation compared to ISM generated RIRs which may cause signal clipping. Our finding is that ISM-generated RIRs often lead to over-excitation. As the IHM data contains strong audio signals, in combination with the amplifying ISM RIR, the augmented training data is heavily clipped. This leads to overall ASR system degradation. ESS RIRs do not have this issue.

TABLE VI
COMPARISON OF THE EFFECT OF PASSIVATION **PAS** – TOP PANEL, AND DELAY SHIFT **SHI** – BOTTOM PANEL, ON VARIOUS RIR SETS

System	WER [%]			
	IHM		SDM	
	dev	eval	dev	eval
ihm.AR2k.vis.ct3m.perfile	25.80	28.40	44.38	48.48
ihm.AR2k.**pas**.vis.ct3m.perfile	21.72	23.24	43.42	48.46
ihm.RR306.vis.ct3m.perfile	25.83	28.35	44.05	48.55
ihm.RR306.**pas**.vis.ct3m.perfile	25.44	28.42	44.14	48.54
ihm.AR306.pas.vis.ct3m.per1seg	21.40	23.00	41.72	46.76
ihm.AR306.pas.**shi**.vis.ct3m.per1seg	21.46	23.35	41.78	47.30
ihm.RR306.pas.vis.ct3m.per1seg	25.22	27.18	43.26	47.42
ihm.RR306.pas.**shi**.vis.ct3m.per1seg	25.32	27.72	43.10	47.12
ihm.RR30.pas.vis.ct2m.f2f.per1seg	23.12	24.80	42.30	46.36
ihm.RR30.pas.**shi**.vis.ct2m.f2f.per1seg	22.88	24.42	42.42	46.44

The last four lines aim at RIR delay compensation. When analyzing the distribution of delays, we found, that artificial RIRs have a peak at 0 seconds with about $1/4$ of them uniformly distributed from 0 to 0.02 seconds (2 frames). On the other hand, real RIRs delay distribution is Gaussian with peak at 0 and tailing to ± 0.05 seconds with extreme values reaching 0.14 second (14 frames). A negative delay can be caused, for example, by less precise loudspeaker to microphone distance measurement. A small positive delay is not so substantial as it only leads to delaying the reverberated audio with respect to the alignment, and a delay within 1–2 frames can be considered as wanted variability. Larger delays may cause degradation due to desynchronization of the audio and alignment in NN training (see Section VI-B). However, a negative delay is critical, as when we try to compensate it, the beginning of RIR (containing the important direct sound and early reflections!) is cut off. Such trimmed RIR is damaged, as it does not carry full information on the acoustic environment anymore.

The results (lines 5 and 6 in Table VI) show that applying delay compensation (synchronizing all RIRs to start at 0 seconds) for artificial RIRs does not have significant impact except for small deterioration for SDM eval set. Applying delay compensation for real RIRs (lines 7 to 10 in Table VI) has mixed results. Small errors in distance measurement can actually bring wanted variability to the augmented data in some cases. We decided to use passivation but not delay compensation in further experiments, as the former has clearly gain, but the results of the later can be considered as statistical noise.

F. Simulated Versus Real Room Impulse Responses on AMI Data

We compared the influence of artificial RIRs (ISM generated) with real RIRs (estimated from BUT ReverbDB using ESS method) in the following experiments. It should be remembered, that the scenario is training ASR to target an unseen environment (AMI meeting rooms) without having any target data. We tried to answer the following questions:

- How many RIRs are sufficient?
- Are artificial RIRs superior to real ones?

TABLE VII
COMPARISON OF VARIOUS ASRS TRAINED ON AUGMENTED IHM TO THE BASELINE (ASR TRAIN ON CLEAN IHM) AND "CHEATING" TARGET (ASR TRAINED ON SDM) SYSTEMS. THE BOTTOM PART COMPARES SYSTEM COMBINATION (ON TRAINING DATA LEVEL)

System	WER [%]			
	IHM		SDM	
	dev	eval	dev	eval
ihm (**baseline**)	20.02	20.04	60.12	72.70
ihm.AR2k.pas.vis.ct3m.per1seg	21.44	23.02	41.80	47.06
ihm.RR306.pas.vis.ct3m.per1seg	25.22	27.18	43.26	47.42
ihm.AR306.pas.vis.ct3m.per1seg	**21.40**	**23.00**	**41.72**	46.76
ihm.RR30.pas.vis.ct2m.f2f.per1seg	23.12	24.80	42.30	**46.36**
ihm.AR30.pas.vis.ct2m.f2f.per1seg	21.86	23.70	41.92	46.76
ihm.RR306.pas.vis.ct3m.per1seg + ihm.AR306.pas.vis.ct3m.per1seg	22.30	23.90	41.80	46.24
ihm.RR30.pas.vis.ct2m.f2f.per1seg + ihm.AR2k.pas.vis.ct3m.per1seg	21.86	23.22	**41.54**	**46.12**
sdm1 (**target**)	29.38	36.74	35.72	39.65

- Are artificial and real RIRs complementary?

We summarized our results in Table VII. The baseline system **ihm** is trained on AMI IHM data using default Kaldi recipe (Section VI-A). This system performs well on in-domain IHM dev (20.02% WER) and eval (20.04% WER) data, but very badly on target (out-of-domain) SDM dev (60.12% WER) and eval (72.70% WER). To have an idea of the best reachable WER, we trained the system on target data – **sdm1**, in the same way as the **ihm**. We achieved expected huge improvement on (in this case in-domain) SDM dev (35.7%) and eval (39.6%) data, but significant deterioration on (now out-of-domain) IHM dev (29.3%) and eval (36.7%) data.

We then applied various data augmentation techniques on IHM training data to simulate the target environment and to achieve an ASR adapted to SDM data, without seeing any SDM data. We use the following notation:

- **RR30** – set of 30 real RIRs including 30 microphones from 4 rooms of BUT Reverb DB (see Section IV) with microphones in a range of 1-2 meters from the loudspeaker and face-to-face orientation.
- **RR306** – set of 306 real RIRs including 306 microphones from the 4 rooms with microphones in a range of $1-3$ meters from the loudspeaker and direct visibility. *RR306* is superset of *RR30*.
- **AR30** – set of 30 artificially generated RIRs with microphones in a range of $1-2$ meters from the loudspeaker and face-to-face orientation. This set is a random draw from a larger set of artificial RIRs with parameters set to as close as possible to the 4 rooms. This set should be comparable to *RR30*.
- **AR306** – set of 306 artificially generated RIRs with microphones in a range of $1-3$ meters from the loudspeaker and direct visibility. This set is a random draw from *AR2k* set
- **AR2k** – set of 2000 artificially generated RIRs with microphones in a range of $1-3$ meters from the loudspeaker and direct visibility. Parameters of the RIRs are as close as possible to the 4 rooms.

By comparing the results of five systems from the upper part of Table VII, we can conclude that using a larger set of RIRs is not always beneficial — see the significant gain when going from *AR30* to *AR306*, but no gain or even deterioration when going from *AR306* to *AR2k* and a significant deterioration for real RIRs – going from *RR30* to *RR306*. We conclude that a careful selection of RIRs covering the target scenario is important.

Comparing the artificial RIR (*AR**) to real RIR (*RR**) systems shows no clear winner. Artificial RIRs have a significant advantage in working well on IHM data too, making the ASR more robust on both IHM and SDM data. On the other hand, *ihm.RR30.pas.vis.ct2m.f2f.per1seg* system is significantly better on the SDM eval set.

Finally, artificial and real RIRs seem to be complementary and their combination is beneficial (see bottom part of Table VII). The combination was done on the level of training data by taking one half of data augmented by artificial RIRs and one half of data augmented by real RIRs (in order to always train on the same amount of data). *RR30 + AR2k* achieved the best WER on the SDM data set with small deterioration on IHM data set compared to the best single systems.

VII. CONCLUSION AND FUTURE WORK

This paper presents BUT ReverbDB, a public set of RIRs, noise and retransmitted data for ASR and SRE development and testing. The set is available for free under a non-restrictive CC-BY license, and covers non-standard positions of microphones that are interesting for investigation/intelligence scenarios.Currently, the set contains data from 8 rooms and will continue to grow. We believe that our paper can serve as a cook-book of how to collect such dataset.

A set of experiments aiming at the ASR test data processing was performed in order to check and validate the database, with interesting findings: Clock asynchronicity problem in RIR estimation by MLS technique was studied and we found that it can be fixed by estimating the clock ratio using cross-correlation (when applied, we obtained comparable WER results as with the ESS technique). We also confirmed other papers' conclusion on the importance of adding real noise in ASR test data preparation. Finally, we observed a clear superiority of real RIRs over artificial ones.

ASR training data augmentation experiments targeted training of an ASR system on data augmented by real or artificial RIRs. We have found the passivation of RIR is extremely important, and recommend checking this issue in other RIR datasets. We also concluded that knowing the target room configuration is beneficial, as we obtained better results with a few carefully selected RIRs than with a huge number of randomly picked ones. In real applications, this calls for a system capable of extracting RIRs from reverberated audio and its use for the augmentation of training data. We have also shown that real and artificial RIRs are complementary, and investigated into a number of technical (but nonetheless important) issues such as reverberation of long audio files per speaker, and RIR delay compensation.

In future work, we would like to grow our data-set, and extend it with real speech data. Our experimental work will include

investigation into the influence of having just one or two IRs from one room rather than many IRs from one room, a simplification of ASR system (i.e. producing results without i-Vector adaptation) and also changing the noise within each speaker adaptation segment.

APPENDIX A
MEASURING RIRS IN BUT REVERBDB

This section contains a more detailed description of the hardware and software used and metadata collected. Even more details accompanied with photos are available in the technical report which is part of the BUT ReverbDB release.[13]

A. Hardware

1) Audio Recording: Our requirements on recorded audio are a large amount of channels in high quality and sample-to-sample synchronization across all channels (see Section III) at reasonable price.[14] We decided to design our own hardware with the help of colleagues from Audified.[15] The device is based on Analog Devices development board SC589 equipped with an ARM Cortex A5 processor and Sharc DSP processor. The processor board is connected to two 16-channel boards equipped with 96 kHz, 24bit, AKM A/D converters with software driven gains and phantom power. The sampled audio data are assembled in TCP/IP packets (interleaving format with timestamps) and sent through Ethernet to a local recording device. The 32 channels here are reconstructed and stored on a hard-drive as 32 PCM audio files. Any packet drop-outs are reported into a log file.

2) Audio Playback: We used an external USB stereo soundcard with symmetrical outputs. We played our audio data in the left channel together with a control signal played in the right channel. The control signal allows us to detect possible problems (caused by a playback buffer under-run, samples drop, packet loss, etc.) and to split the long raw recordings back to the retransmitted audio corpus (parallel corpus).

The control signal is recorded as channel 32 on the recording device. The left channel is fed to the loudspeaker – Adam audio A7X studio monitor[16]. The loudspeaker is placed on a wheeled stand with a settable height. The loudspeaker is placed in the following positions in each room:

- Sitting person: Usually in front of a computer monitor or a table simulating a sitting person (about 140cm above the floor).
- Standing person: Placed randomly in the room where a person can stay (about 170 cm above the floor).
- Noise source: Simulates position of a source of noise, for example a radio, air-condition (AC), fan, etc. The reasoning is to collect RIR of noise source and then generate "real" noises by, for example, reverberating an FM radio audio stream using this RIR.
- Non-standard position: Directed to the ceiling, or floor, lying on the floor, etc.

3) Noise Sources: Most environments are without any additional noise source. The real noises include AC, vents, or common street noise coming through windows. We added artificial noise sources in a few recording sessions. This is marked in the meta-data. We used a Tecsun PL-680 radio receiver tuned to a random local FM station as another source of noise. This noise source is placed in the usual radio positions in the room.

B. Microphones, Mountings, and Positions

We use two types of microphone capsules, both with symmetrical wiring and phantom-powered:

- *standard microphone capsule* (a majority of our microphones) includes an omnidirectional electret condenser microphone module – `PMOF-6027PN-42UQ`.
- *Sennheiser* MKE 2 omnidirectional microphone.

They are placed in several mountings:

1) Spherical Array Mounting: In order to cover the microphone array use-case, we designed a spherical 8-channel array. It consists of 8 *standard microphone capsules* placed in an 8 cm diameter sphere on two parallel planes (4 per each). Microphones are placed in square vertices. The two vertices are rotated by 45°. The orientation of the microphones is from the sphere center. This microphone mounting is usually placed where a similar device (a smart home assistant) is expected in the room.

2) Internet-of-Things Mounting: We mounted two standard microphone capsules into plastic boxes with magnets glued on. These devices are usually attached to a wall or some metal object mounted on a wall.

3) Stand Mounting: 6 to 10 microphones are mounted on a stand. These are then placed on floor, table, etc. and adjusted to desired microphone position and direction. Some of the microphones are also mounted to a computer monitor, lamp, and other objects simulating table-top microphones.

4) Laid Mounting: 5 to 10 microphones are just laid on a chair, table, cupboard, shelf, etc. The microphone is usually oriented approximately towards the sound source.

5) Hidden Laid Mounting: Some of the laid microphones are partly or fully hidden in an object (occluded microphones). This simulates the placing of "bugs" and listening devices. The place is described in the particular microphone placement meta-data. We hid the microphone in a shelf, drawer, waste bin, flower, vent or behind painting, white board, etc.

6) In-Air Mounting: About 5 microphones are placed in the air close to the ceiling. We use fishing rods here to place the microphones in the upper corners, close to various sensors (smoke detectors), lights, projectors, etc. We also let one or two microphones just hang down and be in the space far from any obstacles.

C. Meta-Data

We generated a lot of meta-data to provide details on: 1) the room (environment), 2) loudspeaker placing(s), 3) microphones placings. The meta-data is available in the text files. We provide several coordinate systems allowing for easy work with our

[13] https://speech.fit.vutbr.cz/software/but-speech-fit-reverb-database

[14] We are aware of AVS or other hi-end solutions with master/slave clock bus etc., but these were unacceptable for our budget.

[15] http://www.audified.com

[16] RMS: 100W, Frequency response: 42 Hz – 50 kHz, Crossover frequency: 2.5 Hz, Size: 337 mm × 201 mm × 280 mm, Weight: 9.2 kg, Bass reflex in front. http://www.adam-audio.com/en/pro-audio/products/a7x/description

data set. We use absolute and relative Cartesian (depth, width, height) and spherical (distance, azimuth, elevation) coordinates for microphone and loudspeaker positions. We also use azimuth and elevation for microphone or loudspeaker orientation.

The origin for *relative measurements* is the placement of the loudspeaker (speech source). Microphone to loudspeaker distance can, therefore, be easily obtained by looking to the relative distance of the microphone. In addition to the size of the room, we store photos, description, type, size, temperature, materials, amount of furniture, and background noise level.

We can place several microphone setups in every room, however, we use mainly just one microphone setup per room. On the other hand, we usually place the loudspeaker(s) in several positions for every microphone setup. We try to have at least five distinct positions here. The first position of the loudspeaker is the one we use for measuring the coordinates of all microphones and their meta-data. One loudspeaker setup can consists of one or more physical loudspeakers. The first loudspeaker is always the one playing the audio (speech data, sine sweeps, MLS, etc.). The others may be used as noise sources (radio in the background, etc.). We store coordinates (position), orientation (facing), and the type of the loudspeaker as meta-data.

ACKNOWLEDGMENT

The authors would like to thank Kamil Chalupníček and Ondřej Novotný for helping us to collect the BUT ReverbDB.

REFERENCES

[1] M. Karafiát *et al.*, "Training data augmentation and data selection," *New Era for Robust Speech Recognition: Exploiting Deep Learning*. Heidelberg, Germany: Springer International Publishing, 2017, pp. 245–260.

[2] J. Melot, N. Malyska, J. Ray, and W. Shen, "Analysis of factors affecting system performance in the ASpIRE challenge," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Dec. 2015, pp. 512–517.

[3] L. Mošner, O. Plchot, P. Matějka, O. Novotný, and J. Černocký, "Dereverberation and beamforming in robust far-field speaker recognition," in *Proc. Interspeech*. International Speech Communication Association, 2018, pp. 1334–1338.

[4] L. Mošner, P. Matějka, O. Novotný, and J. Černocký, "Dereverberation and beamforming in far-field speaker recognition," in *Proc. ICASSP*. IEEE Signal Processing Society, 2018, pp. 5254–5258.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210.

[6] C. Greenberg, A. Martin, D. Graff, L. Brandschain, and K. Walker, *2010 NIST Speaker Recognition Evaluation Test Set LDC2017S06*, Linguistic Data Consortium, Philadelphia, PA, USA, 2017.

[7] T. Yoshioka *et al.*, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[8] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Dec. 2015, pp. 547–554.

[9] A. Janin *et al.*, "The ICSI Meeting Corpus," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, 2003, pp. 364–367.

[10] S. H. K. Parthasarathi, S. Chang, J. Cohen, N. Morgan, and S. Wegmann, "The blame game in meeting room ASR: An analysis of feature versus model errors in noisy and mismatched conditions," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, May 2013, pp. 6758–6762.

[11] T. Hain *et al.*, "The AMI system for the transcription of speech in meetings," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing – ICASSP '07*, Apr. 2007, vol. 4, pp. IV-357–IV-360.

[12] 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 Automatic Speech Recognition Understanding*, Nov. 2005, pp. 357–362.

[13] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Mixer 6," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[14] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, Mar. 2017, pp. 5220–5224.

[15] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL "CHiME" speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 27, pp. 621–633, May 2013.

[16] E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, Dec. 2013, pp. 162–167.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third "CHiME" speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Dec. 2015, pp. 504–511.

[18] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth "CHiME" speech separation and recognition challenge: Dataset, task and baselines," Tech. Rep., 2018. [Online]. Available: https://arxiv.org/pdf/1803.10609.pdf

[19] K. Kinoshita *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics*, Oct. 2013, pp. 1–4.

[20] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Advances Signal Process.*, vol. 2016, no. 1, p. 7, Jan. 2016.

[21] M. Ravanelli, P. Svaizer, and M. Omologo, "Realistic multi-microphone data simulation for distant speech recognition," Nov. 2017, arXiv:1711.09470v1.

[22] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digital Signal Processing*, Jul. 2009, pp. 1–5.

[23] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Acoustic characterization of environments (ACE) challenge results technical report," Tech. Rep., Imperial College London, 2017. [Online]. Available: https://arxiv.org/pdf/1606.03365.pdf

[24] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio Speech Lang. Proc.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[25] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. Language Resources Evaluation Conf.*, 2000, pp. 356–360.

[26] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Dec. 2015, pp. 275–282.

[27] N. Bertin *et al.*, "A French Corpus for distant-microphone speech processing in real homes," in *Proc. Annu. Conf. Int. Speech Communication Assoc., INTERSPEECH*, San Francisco, NC, USA, Sep. 2016, pp. 2781–2785.

[28] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The sweet-home speech and multimodal corpus for home automation interaction," in *Proc. 9th Edition Language Resources Evaluation Conf. (LREC)*, Reykjavik, Iceland, 2014, pp. 4499–4506. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00953006

[29] M. Ravanelli and M. Omologo, "On the selection of the impulse responses for distant-speech recognition based on contaminated speech training," in *Proc. Annu. Conf. Int. Speech Communication Assoc., INTERSPEECH*, Sep. 2014, p. 4.

[30] M. R. Schroeder, "Integrated-impulse method for measuring sound decay without using impulses," *J. Acoust. Soc. Am.*, vol. 66, pp. 497–500, 1979.

[31] C. Dunn and M. J. Hawksford, "Distortion immunity of MLS-derived impulse response measurements," *J. Audio Eng. Soc.*, vol. 41, no. 5, pp. 314–335, 1993.

[32] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1484–1488, 1981.

[33] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Eng. Soc. Conv.*, Feb. 2000. [Online]. Available: http://www.aes.org/elib/browse.cfm?elib=10211

[34] G. B. Stan, J. J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *J. Audio Eng. Soc.*, vol. 50, pp. 249–262, 2002.

[35] S. W. Golomb, *Shift Register Sequences*. Laguna Hills, CA, USA: Aegean Park Press, 1981.

[36] J. Vanderkooy, "Aspects of MLS measuring systems," *J. Audio Eng. Soc.*, vol. 42, no. 4, pp. 219–231, 1994. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=6951

[37] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *Proc. 20th Eur. Signal Processing Conf. (EUSIPCO)*, Aug. 2012, pp. 1668–1672.

[38] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Proc. Audio Eng. Soc. Conv.*, May 2007, pp. 1–21. [Online]. Available: http://www.aes.org/elib/browse.cfm?elib=14106

[39] M. R. P. Thomas, "MLS project," Tech. Rep., 2009. [Online]. Available: http://www.commsp.ee.ic.ac.uk/mrt102/projects/mls.html

[40] S. Siltanen, T. Lokki, and L. Savioja, "Rays or waves? Understanding the strengths and weaknesses of computational room acoustics modeling techniques," in *Proc. Int. Symp. Room Acoustics (ISRA2010)*, Melbourne, Australia, Aug. 29–31, 2010.

[41] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 708–730, 2015.

[42] R. D. Ciskowski and C. A. Brebbia, *Boundary Element Methods in Acoustics*. New York: Elsevier Applied Science, 1991.

[43] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*. New York: Springer, c1998.

[44] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[45] A. Krokstad, S. Strom, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. Sound Vib.*, vol. 8, no. 1, pp. 118–125, 1968.

[46] C. Kim *et al.*, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *Proc. Annu. Conf. Int. Speech Communication Assoc., INTERSPEECH*, Aug. 2017, pp. 379–383.

[47] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Sep. 2010. [Online]. Available: https://github.com/ehabets/RIR-Generator

[48] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, April 2018*. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers Inc., Sep. 2018, pp. 351–355.

[49] O. Glembek, M. Karafiát, L. Burget, and J. Černocký, "Czech speech recognizer for multiple environments," in *Proc. Radioeletronika*, 2006, pp. 1–4.

[50] F. Grézl, M. Karafiát, and K. Veselý, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. ICASSP*. IEEE Signal Processing Society, 2014, pp. 7704–7708.

[51] H. Kuttruff, *Room Acoustics*, 5th ed. Boca Raton, FL, USA: CRC Press, Jun. 26, 2009.

[52] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoustical Soc. Am.*, vol. 37, no. 3, pp. 409–412, 1965.

[53] A. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, vol. 34. Melville, NY, USA: ASA, Jun. 1989.

[54] I. McCowan *et al.*, "The AMI Meeting Corpus," in *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, L. Noldus, F. Grieco, L. Loijens, and P. Zimmerman, Eds. Wageningen, The Netherlands: Noldus Inf. Technol., Aug. 2005, pp. 137–140.

[55] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*. IEEE Signal Processing Society, Dec. 2011.

[56] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *Proc. INTERSPEECH*. ISCA, 2015, pp. 2440–2444.

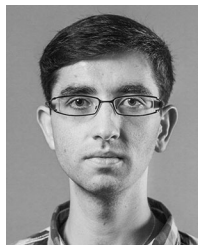
Igor Szöke (M'04) received the Ph.D. degree in information technology from the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic, in 2010.

He is currently an Assistant Professor with the Faculty of Information Technology (FIT), Brno University of Technology (BUT). Since 2003, he has been with the BUT Speech@FIT research group. His research interests include machine learning and speech data mining (automatic speech recognition, spoken term detection, and data augmentation). He co-founded Phonexia in 2006 and ReplayWell in 2011.

Miroslav Skácel received the master's degree in information technology in 2015 from the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic where he is currently working towards the Ph.D. degree.

His research interests include low-resource automatic speech recognition, speech signal enhancement, and deep learning.

He has been a member of BUT Speech@FIT research group since 2012.



Ladislav Mošner received the master's degree in information technology from the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic, in 2017. He is working towards the Ph.D. degree at the BUT Speech@FIT research group.

His interests include multichannel processing (beamforming), noise and reverberation robustness, and speech enhancement and discriminative training in speaker recognition.

Jakub Paliesek received the bachelor's degree from Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic, in 2018, after submitting the thesis "Measurement of Environment Acoustics Impact on Speech Recognition Accuracy." He is now a master's degree student at the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic.

Jan (Honza) Černocký (M'01–SM'08) received the Ing. (MSc.) degree from the Brno University of Technology (BUT), Brno, Czechia, in 1993, and the Doctorate degree from Universite Paris XI Orsay, Orsay, France, and BUT, in 1998.

He is currently an Associate Professor and Head of the Department of Computer Graphics and Multimedia, Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic. He also serves as Managing Director of BUT Speech@FIT research group. His research interests include artificial intelligence, signal processing, and speech data mining (speech, speaker, and language recognition). He is responsible for signal and speech processing courses at FIT BUT. In 2006, he co-founded Phonexia. He is general chair of Interspeech 2021 in Brno.