# Chapter 15
# The REVERB Challenge: A Benchmark Task for Reverberation-Robust ASR Techniques

**Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A.P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka**

**Abstract** The REVERB challenge is a benchmark task designed to evaluate reverberation-robust automatic speech recognition techniques under various conditions. A particular novelty of the REVERB challenge database is that it comprises both real reverberant speech recordings and simulated reverberant speech, both of which include tasks to evaluate techniques for 1-, 2-, and 8-microphone situations. In this chapter, we describe the problem of reverberation and characteristics of the REVERB challenge data, and finally briefly introduce some results and findings useful for reverberant speech processing in the current deep-neural-network era.

K. Kinoshita (✉) • M. Delcroix • T. Nakatani • T. Yoshioka
NTT Communication Science Laboratories, NTT Corporation, 2-4, Hikaridai, Seika-cho, Kyoto, Japan
e-mail: kinoshita.k@lab.ntt.co.jp

S. Gannot
Bar-Ilan University, Ramat Gan, Israel

E.A.P. Habets
International Audio Laboratories Erlangen, Erlangen, Germany

R. Haeb-Umbach
University of Paderborn, Paderborn, Germany

W. Kellermann • R. Maas
Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany

V. Leutnant
Amazon Development Center Germany GmbH, Aachen, Germany

B. Raj
Carnegie Mellon University, Pittsburgh, PA, USA

A. Sehr
Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany

S. Watanabe et al. (eds.), *New Era for Robust Speech Recognition*,
DOI 10.1007/978-3-319-64680-0_15

## 15.1 Introduction

Speech signal-processing technologies have advanced significantly in the last few decades, and now play various important roles in our daily lives. Especially, speech recognition technology has advanced rapidly, and is increasingly coming into practical use, enabling a wide spectrum of innovative and exciting voice-driven applications. However, most applications consider a microphone located near the talker as a prerequisite for reliable performance, which prohibits further progress in automatic speech recognition (ASR) applications.

Speech signals captured with distant microphones inevitably contain interfering noise and reverberation, which severely degrade the speech intelligibility of the captured signals [16] and the performance of ASR systems [4, 18]. A noisy reverberant observed speech signal $y(t)$ at time $t$ can be expressed as

$$y(t) = h(t) * s(t) + n(t), \tag{15.1}$$

where $h(t)$ corresponds to the room impulse response between the speaker and the microphone, $s(t)$ to the clean speech signal, $n(t)$ to the background noise, and $*$ to the convolution operator. Note that the primary focus of interest in the REVERB challenge is on reverberation, i.e., the effect of $h(t)$ on $s(t)$, and techniques which address it.

Research on reverberant speech processing has made significant progress in recent years [11, 19], mainly driven by multidisciplinary approaches that combine ideas from room acoustics, optimal filtering, machine learning, speech modeling, enhancement, and recognition. The motivation behind the REVERB challenge was to provide a common evaluation framework, i.e., tasks and databases, to assess and collectively compare algorithms and gain new insights regarding the potential future research directions for reverberant speech-processing technology.

This chapter summarizes the REVERB challenge, which took place in 2014 as a community-wide evaluation campaign for speech enhancement (SE) and ASR techniques [6, 7, 13]. While other benchmark tasks and challenges [1, 12, 17] mainly focus on the noise-robustness issue and sometimes only on a single-channel scenario, the REVERB challenge was designed to test robustness against *reverberation* under moderately noisy environments. The evaluation data of the challenge contains both *single-channel* and *multichannel* recordings, both of which comprise *real recordings* and *simulated data*, which has similar characteristics to real recordings. Although the REVERB challenge contains two tasks, namely SE and ASR tasks, we focus only on the latter task in this chapter.

The remainder of this chapter is organized as follows. In Sect. 15.2, we describe the scenario assumed in the challenge and details of the challenge data. Section 15.3 introduces results for baseline systems and top-performing systems. Section 15.4 provides a summary of the chapter and potential research directions to further develop reverberation-robust ASR techniques.

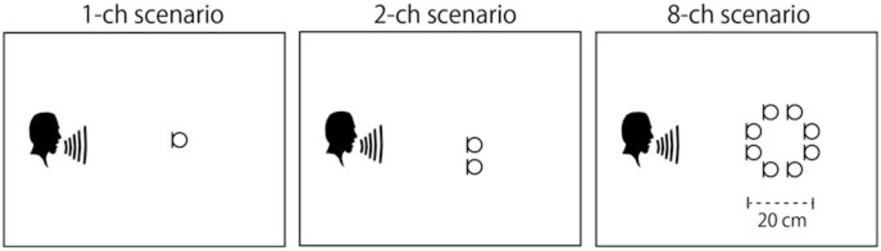

**Fig. 15.1** Scenarios assumed in the REVERB challenge

## 15.2 Challenge Scenarios, Data, and Regulations

### *15.2.1 Scenarios Assumed in the Challenge*

Figure 15.1 shows the three scenarios considered in this challenge [6, 7], in which an utterance spoken by a single spatially stationary speaker is captured with single-channel (1-ch), two-channel (2-ch), or eight-channel (8-ch) circular microphone arrays in a moderately noisy reverberant room. In practice, we commonly encounter this kind of acoustic situation when, e.g., we attend a presentation given in a small lecture room or a meeting room. In fact, the real recordings used in the challenge were recorded in an actual university meeting room, closely simulating the acoustic conditions of a lecture hall [10]. The 1-ch and 2-ch data are simply a subset of the 8-ch circular-microphone-array data. The 1-ch data were generated by randomly picking up one of eight microphones, while the 2-ch data were generated by randomly picking up adjacent two microphones from the eight microphones. For more details of the recording setting, refer to a document in the "download" section of the challenge webpage [13].

### *15.2.2 Data*

For the challenge, the organizers provided a dataset which consisted of training data and test data. The test data comprised a development (Dev) test set and an evaluation (Eval) test set. All the data was provided as 1-ch, 2-ch, and 8-ch reverberant speech recordings at a sampling frequency of 16 kHz, and is available through the challenge webpage [13] via its "download" section. An overview of all the datasets is given in Fig. 15.2. Details of the test and training data are given in the following subsections.

#### 15.2.2.1 Test Data: Dev and Eval Test Sets

By having the test data (i.e., the Dev and Eval test sets) consisting of both real recordings (RealData) and simulated data (SimData), the REVERB challenge

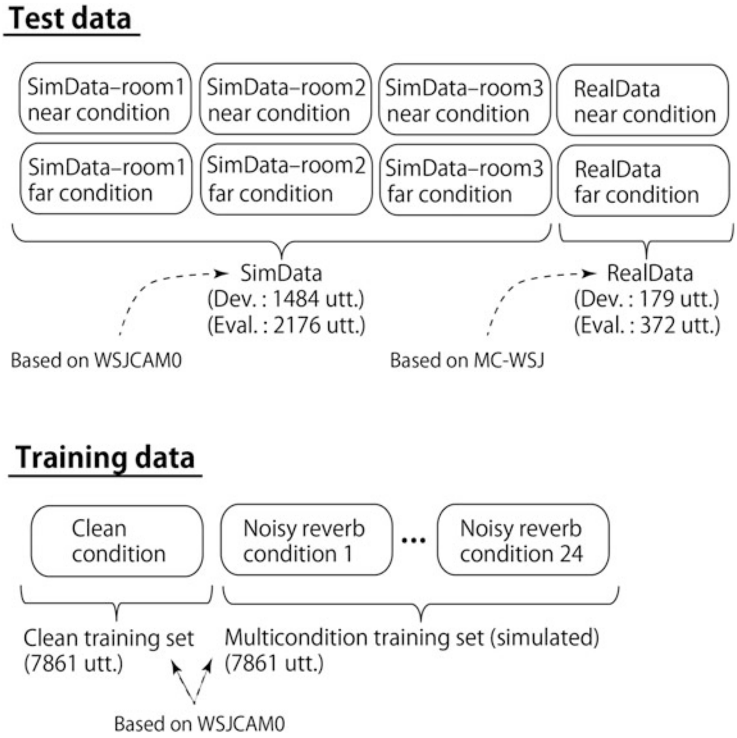
Test data
SimData–room1 near condition
SimData–room2 near condition
SimData–room3 near condition
RealData near condition
SimData–room1 far condition
SimData–room2 far condition
SimData–room3 far condition
RealData far condition
SimData (Dev. : 1484 utt.) (Eval. : 2176 utt.)
RealData (Dev. : 179 utt.) (Eval. : 372 utt.)
Based on WSJCAM0
Based on MC-WSJ
Training data
Clean condition
Noisy reverb condition 1
...
Noisy reverb condition 24
Clean training set (7861 utt.)
Multicondition training set (simulated) (7861 utt.)
Based on WSJCAM0

**Fig. 15.2** Overview of datasets used in the REVERB challenge

provided researchers with an opportunity to thoroughly evaluate their algorithms for (1) practicality in realistic conditions and (2) robustness against a wide range of reverberant conditions.

- *SimData* is composed of reverberant utterances generated based on the WSJCAM0 British English corpus [9, 14]. These utterances were artificially distorted by convolving clean WSJCAM0 signals with measured room impulse responses (RIRs) and subsequently adding measured stationary ambient noise signals with a signal-to-noise ratio (SNR) of 20 dB. SimData simulated six different reverberation conditions: three rooms with different volumes (small, medium, and large) and two distances between a speaker and a microphone array (near = 50 cm and far = 200 cm). Hereafter, the rooms are referred to as SimData-room1, -room2, and -room3. The reverberation times (i.e., T60) of SimData-room1, -room2, and -room3 were about 0.3, 0.6, and 0.7 s, respectively. The RIRs and added noise were recorded in the corresponding reverberant room with an 8-ch circular array with a diameter of 20 cm. The recorded noise was stationary diffuse background noise, which was mainly caused by the air conditioning systems in the rooms, and thus has relatively large energy at lower frequencies.

- *RealData*, which comprises utterances from the MC-WSJ-AV British English corpus [8, 10], consists of utterances spoken by human speakers in a noisy and reverberant meeting room. RealData contains two reverberation conditions: one room and two distances between the speaker and the microphone array (near, i.e., about 100 cm, and far, i.e., about 250 cm). The reverberation time of the room was about 0.7 s [10]. Judging by the reverberation time and the distance between the microphone array and the speaker, the characteristics of RealData resemble those of the SimData-room-3-far condition. The text prompts for the utterances used in RealData and in part of SimData were the same. Therefore, we can use the same language and acoustic models for both SimData and RealData. For RealData recordings, a microphone array which had the same array geometry as the one used for SimData was employed.

For both SimData and RealData, we assumed that the speakers stayed in the same room for each test condition. However, within each condition, the relative speaker–microphone position changed from utterance to utterance. The term "test condition" in this chapter refers to one of the eight reverberation conditions that comprise two conditions in RealData and six conditions in SimData (see Fig. 15.2).

#### 15.2.2.2 Training Data

As shown in Fig. 15.2, the training data consisted of (1) a clean training set taken from the original WSJCAM0 training set and (2) a multicondition (MC) training set. The MC training set was generated from the clean WSJCAM0 training data by convolving the clean utterances with 24 measured room impulse responses and adding recorded background noise at an SNR of 20 dB. The reverberation times of the 24 measured impulse responses for this dataset range roughly from 0.2 to 0.8 s. Different recording rooms were used for the test data and the training data, while the same set of microphone arrays was used for the training data and SimData.

### *15.2.3 Regulations*

The ASR task in the REVERB challenge was to recognize each noisy reverberant test utterance without a priori information about the speaker identity/label, room parameters such as the reverberation time, the speaker–microphone distance and the speaker location, and the correct transcription. Therefore, systems had to perform recognition without knowing which speaker was talking in which acoustic condition.

Although the relative speaker–microphone position changed randomly from utterance to utterance, it was allowed to use all the utterances from a single test condition and to perform full-batch processing such as environmental adaptation of the acoustic model (AM). This regulation was imposed to focus mainly on the effect of environmental adaptation rather than speaker adaptation.

## 15.3 Performance of Baseline and Top-Performing Systems

To give a rough idea of the degree of difficulty of the REVERB challenge data, this section summarizes the performance achieved by the baseline and some notable top-performing systems.

### *15.3.1 Benchmark Results with GMM-HMM and DNN-HMM Systems*

First of all, let us introduce the performance that can be obtained with Gaussian-mixture-model–hidden-Markov-model (GMM-HMM) recognizers and deep-neural-network–hidden-Markov-model (DNN-HMM) recognizers without front-end processing. Table 15.1 shows the results of two versions of Kaldi-based baseline GMM-HMM recognizers [5], and two versions of simple DNN-HMM recognizers which were prepared by two different research institutes independently [2, 3]. The table shows that even a very complex GMM-HMM system (the second system in the table) is outperformed by simple fully connected DNN-HMM systems for both SimData and RealData, which clearly indicates the superiority of the DNN-based AM over the GMM-based AM. However, although these improvements are notable and may support a claim that DNNs are robust in adverse environments, the achieved performances are actually still very far from the WERs obtained with clean speech, which correspond to 3.5% for SimData and 6.1% for RealData. The goal of reverberation-robust ASR techniques is to close the performance gap between clean speech recognition and reverberant speech recognition. Note that the big gap between the SimData and RealData performance in Table 15.1 is partly due to the fact that many of the SimData settings were less reverberant than the RealData setting.

**Table 15.1** Word error rate (WER) obtained by baseline GMM-HMM systems and simple DNN systems without front-end processing (Eval set) (%)

| System | SimData WER (%) | RealData WER (%) |
|---|---|---|
| Baseline multicondition GMM-HMM system with bigram LM [5] | 28.8 | 54.1 |
| Baseline multicondition GMM-HMM system with MMI AM training, trigram LM, fMLLR, MBR decoding [5] | 12.2 | 30.9 |
| DNN system with fully connected seven hidden layers [2] with trigram LM | 8.6 | 28.5 |
| DNN system with fully connected five hidden layers [3] with trigram LM | 8.9 | 28.2 |

*LM* language model

## 15.3.2 Top-Performing 1-ch and 8-ch Systems

Next, let us introduce the performance of the top-performing 1-ch systems to show how they are achieving their goal currently. In this subsection, for the sake of simplicity, we present only the results from utterance-based batch processing systems, which usually are suitable for online ASR applications. In addition, the results in this subsection are based on the baseline multicondition training dataset and the conventional trigram LM, excluding the effect of data augmentation and advanced techniques for LM. While there are a number of systems proposed to improve 1-ch ASR performance, among them, the systems proposed by [2, 3, 15] achieved good performances as shown in Table 15.2. Delcroix et al. [2] achieved 7.7% for SimData and 25.2% for RealData by employing linear-prediction-based dereverberation (introduced in Chap. 2) and a simple DNN-based AM. On the other hand, Giri et al. [3] achieved similar performance, i.e., 7.7% for SimData and 27.5% for RealData, by taking a completely different approach. They employed no front-end enhancement technique, but instead fully extended the capability of a DNN-based AM with multitask learning and an auxiliary input feature representing reverberation time [3]. Tachioka et al. [15] took a rather traditional approach, that is, spectral-subtraction-based dereverberation (introduced in Chap. 20) and system combination based on many GMM-SGMM AM-based systems and DNN AM-based systems, and achieved WERs of 8.5% for SimData and 23.7% for RealData.

Now, let us introduce the performance of the top-performing multichannel (here, 8-ch) systems. Tachioka et al. [15] achieved 6.7% for SimData and 18.6% for RealData by additionally employing an 8-ch delay–sum beamformer on top of their 1-ch system. Delcroix et al. [2] achieved 6.7% for SimData and 15.6% for RealData by employing 8-ch linear-prediction-based dereverberation (introduced in Chap. 2),

**Table 15.2** WER obtained by top-performing 1-ch and 8-ch utterance-based batch processing systems (%)

| System | SimData<br>WER (%) | RealData<br>WER (%) |
|---|---|---|
| *1-ch* | | |
| Linear-prediction-based dereverb<br>+ DNN [2] | 7.7 | 25.2 |
| DNN with multitask learning<br>and auxiliary reverb time information [3] | 7.7 | 27.5 |
| Spectral-subtraction dereverb<br>+ system combination of GMM and DNN recognizers [15] | 8.5 | 23.7 |
| *8-ch* | | |
| Linear-prediction-based dereverb + MVDR beamformer<br>+ DNN [2] | 6.7 | 15.6 |
| Delay–sum beamformer + spectral-subtraction dereverb<br>+ system combination of GMM and DNN recognizers [15] | 6.7 | 18.6 |

an 8-ch minimum variance distortionless response (MVDR) beamformer, and a simple DNN-based AM. These results clearly show superiority and importance of multichannel linear-filtering-based enhancement processing. Note that details of the other contributions presented in the REVERB challenge and their effectiveness are summarized in [7].

In summary, based on these results, we confirmed the importance of multichannel linear-filtering-based enhancement, an advanced DNN-based AM, and DNN-related techniques such as auxiliary input features. In addition, it was confirmed that, as in other ASR tasks, system combination provided consistently significant performance gains.

### *15.3.3 Current State-of-the-Art Performance*

Table 15.3 serves as a reference for the current state-of-the-art performance obtained with the challenge data. All of these results were obtained with full-batch processing systems, which usually incorporate environmental adaptation and are generally suitable only for offline ASR applications. The major difference between these results and the ones in Table 15.2 lies in the back-end techniques. Specifically, the systems shown in Table 15.3 employ additionally (a) artificially augmented training data for AM training, (b) full-batch AM adaptation for environmental adaptation, i.e., additional back-propagation training using test data taken from a test condition, and (c) a state-of-the-art LM, i.e., a recurrent neural network (RNN) LM.

**Table 15.3** Current state-of-the-art performance for 1-ch, 2-ch, and 8-ch scenarios (Eval set) (%)

| System | SimData<br>WER (%) | RealData<br>WER (%) |
|---|---|---|
| *1-ch* | | |
| 1-ch linear-prediction-based dereverb<br>+ DNN-based AM + DNN adaptation + data augmentation<br>+ RNN LM [2] | 5.0 | 15.9 |
| *2-ch* | | |
| 2-ch linear-prediction-based dereverb + 2-ch MVDR beamformer<br>+ 1-ch model-based enhancement<br>+ DNN-based AM + DNN adaptation + data augmentation<br>+ RNN LM [2] | 4.4 | 11.9 |
| *8-ch* | | |
| 8-ch linear-prediction-based dereverb + 8-ch MVDR beamformer<br>+ 1-ch model-based enhancement<br>+ DNN-based AM + DNN adaptation + data augmentation<br>+ RNN LM [2] | 4.1 | 9.1 |

## 15.4 Summary and Remaining Challenges for Reverberant Speech Recognition

This chapter introduced the scenario, data, and results for the REVERB challenge, which was a benchmark task carefully designed to evaluate reverberation-robust ASR techniques. As a result of the challenge, it was shown that notable improvement can be achieved by using algorithms such as linear-prediction-based dereverberation and DNN-based acoustic modeling. However, at the same time, it was found that there still remain a number of challenges in the field of reverberant speech recognition. For example the top performance currently obtained for the 1-ch scenario is still very far from that obtained for multichannel scenarios. This is partly due to the fact that there is no 1-ch enhancement technique that can greatly reduce WERs when used with a multicondition DNN AM. Finding a 1-ch enhancement algorithm which works effectively even in the DNN era is one of the key research directions to pursue. It is also important to note that there is still much room for improvement even for multichannel systems, especially for RealData.

The REVERB challenge was a benchmark task to evaluate technologies in reverberant environments where the amount of ambient noise is relatively moderate. However, if the remaining problems mentioned above are resolved in the future by further investigations, we should extend the scenario to include more noise in addition to reverberation, closely simulating more realistic distant speech recognition challenges.

## References

1. Barker, J., Vincent, E., Ma, N., Christensen, C., Green, P.: The PASCAL CHiME speech separation and recognition challenge. Comput. Speech Lang. **27**(3), 621–633 (2013)
2. Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Nobutaka, I., Kinoshita, K., Espi, M., Araki, S., Hori, T., Nakatani, T.: Strategies for distant speech recognition in reverberant environments. Comput. Speech Lang. (2015). doi:10.1186/s13634-015-0245-7
3. Giri, R., Seltzer, M., Droppo, J., Yu, D.: Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5014–5018 (2015)
4. Huang, X., Acero, A., Hong, H.W.: Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall, Upper Suddle River, NJ (2001)
5. Kaldi-based baseline system for REVERB challenge. https://github.com/kaldi-asr/kaldi/tree/master/egs/reverb
6. Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R., Gannot, S., Raj, B.: The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech. In: Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2013)
7. Kinoshita, K., Delcroix, M., Gannot, S., Habets, E., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., Yoshioka, T.: A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP J. Adv. Signal Process. (2016). doi:10.1186/s13634-016-0306-6

8. LDC: Multi-channel WSJ audio. https://catalog.ldc.upenn.edu/LDC2014S03
9. LDC: WSJCAMO Cambridge read news. https://catalog.ldc.upenn.edu/LDC95S24
10. Lincoln, M., McCowan, I., Vepa, J., Maganti, H.K.: The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 357–362 (2005)
11. Naylor, P.A., Gaubitch, N.D.: Speech Dereverberation. Springer, Berlin (2010)
12. Pearce, D., Hirsch, H.G.: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of International Conference on Spoken Language Processing (ICSLP), pp. 29–32 (2000)
13. REVERB Challenge. http://reverb2014.dereverberation.com/
14. Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S.: WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 81–84 (1995)
15. Tachioka, Y., Narita, T., Weninger, F.J., Watanabe, S.: Dual system combination approach for various reverberant environments with dereverberation techniques. In: Proceedings of REVERB Challenge Workshop, p. 1.3 (2014)
16. Tashev, I.: Sound Capture and Processing. Wiley, Hoboken, NJ (2009)
17. Vincent, E., Araki, S., Theis, F.J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B.V., Lutter, D.: The signal separation evaluation campaign (2007–2010): achievements and remaining challenges. Signal Process. **92**, 1928–1936 (2012)
18. Wölfel, M., McDonough, J.: Distant Speech Recognition. Wiley, Hoboken, NJ (2009)
19. Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., Kellermann, W.: Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. **29**(6), 114–126 (2012)