Kinoshita *et al. EURASIP Journal on Advances in Signal Processing* (2016) 2016:7
DOI 10.1186/s13634-016-0306-6

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research

Keisuke Kinoshita[1*], Marc Delcroix[1], Sharon Gannot[2], Emanuël A. P. Habets[3], Reinhold Haeb-Umbach[4], Walter Kellermann[5], Volker Leutnant[6], Roland Maas[5], Tomohiro Nakatani[1], Bhiksha Raj[7], Armin Sehr[8] and Takuya Yoshioka[1]

Abstract

In recent years, substantial progress has been made in the field of reverberant speech signal processing, including both single- and multichannel dereverberation techniques and automatic speech recognition (ASR) techniques that are robust to reverberation. In this paper, we describe the REVERB challenge, which is an evaluation campaign that was designed to evaluate such speech enhancement (SE) and ASR techniques to reveal the state-of-the-art techniques and obtain new insights regarding potential future research directions. Even though most existing benchmark tasks and challenges for distant speech processing focus on the noise robustness issue and sometimes only on a single-channel scenario, a particular novelty of the REVERB challenge is that it is carefully designed to test robustness against *reverberation*, based on *both real, single-channel, and multichannel recordings*. This challenge attracted 27 papers, which represent 25 systems specifically designed for SE purposes and 49 systems specifically designed for ASR purposes. This paper describes the problems dealt within the challenge, provides an overview of the submitted systems, and scrutinizes them to clarify what current processing strategies appear effective in reverberant speech processing.

Keywords: Reverberation, Dereverberation, Automatic speech recognition, Evaluation campaign, REVERB challenge

1 Introduction

Speech signal processing technologies, which have made significant strides in the last few decades, now play various important roles in our daily lives. For example, speech communication technologies such as (mobile) telephones, video-conference systems, and hearing aids are widely available as tools that assist communication between humans. Speech recognition technology, which has recently left research laboratories and is increasingly coming into practical use, now enables a wide spectrum of innovative and exciting voice-driven applications. However, most of these applications consider a microphone located near the talker as a prerequisite for reliable performance, which prevents further proliferation.

Speech signals captured with distant microphones inevitably contain interfering noise and reverberation, which severely degrade the audible speech quality of the captured signals [1] and the performance of automatic speech recognition (ASR) [2, 3]. A reverberant speech signal y(t) at time t can be expressed as

$$y(t) = h(t) * s(t) + n(t), \tag{1}$$

where $h(t)$ corresponds to the room impulse response between the speaker and the microphone, $s(t)$ is the clean speech signal, $n(t)$ is the background noise, and $*$ is the convolution operator.

Although a range of signal processing and speech recognition techniques is available for combating the effect of additive noise (i.e., $n(t)$ in Eq. (1)) [2, 3], finding practical algorithms that can reduce the detrimental effect of reverberation (i.e., $h(t)$ in Eq. (1)) remains one of the toughest challenges in the field of distant-speech enhancement and recognition research.

In recent years, however, research on reverberant speech processing has achieved significant progress in both the audio processing and ASR fields [4, 5], mainly

*Correspondence: kinoshita.k@lab.ntt.co.jp
[1] NTT Communication Science Laboratories, Kyoto, Japan
Full list of author information is available at the end of the article

driven by multidisciplinary approaches that combine ideas from room acoustics, optimal filtering, machine learning, speech modeling, enhancement, and recognition. These novel techniques are now ready to be evaluated for real-world speech enhancement and speech recognition applications.

1.1 Motivation behind REVERB challenge

Numerous papers have reported significant progress on these techniques. However, due to the lack of common evaluation frameworks and databases in this research area, all contributions had different foundations. This complicated accurately determining the importance of the progress that they represent and consequently impedes further technological advancement. Therefore, the motivation behind the challenge is to provide a common evaluation framework, i.e., tasks and databases, to assess and collectively compare the state-of-the-art algorithms and gain new insights regarding the potential future research directions for reverberant speech processing technology.

This paper summarizes the outline and the achievements of the REVERB challenge, which took place in 2014 as a community-wide evaluation campaign for speech enhancement (SE) and ASR techniques handling reverberant speech [6, 7]. Although existing benchmark tasks and challenges [8–10] mainly focus on the noise robustness issue and sometimes only in a single-channel scenario, a particular novelty of the REVERB challenge is that it is carefully designed to test robustness against *reverberation*, based on both *single-channel and multichannel* recordings made under moderately noisy environments. Another novel feature of the challenge is that its entire evaluation is based on *real recordings* and simulated data, part of which has similar characteristics to real recordings. This allows the participants to thoroughly evaluate their algorithms in terms of both the practicality in realistic conditions and robustness against a wide range of reverberant conditions. The challenge is comprised of two types of tasks: ASR and SE. In the ASR task, the submitted systems are evaluated in terms of word error rate (WER), and in the SE task, an SE algorithm's performance is evaluated based on instrumental measures and listening tests evaluating the perceived amount of reverberation and the overall quality of processed signals. The large-scale evaluation of various SE techniques with common instrumental measures and listening tests may provide important insights to help us decide which metrics should be used for properly evaluating SE techniques; this question has not yet been answered satisfactorily.

1.2 Highlight of challenge achievements

The challenge results offer a few important insights for the research community. First, it reveals that notable ASR performance can be accomplished through the careful combination of several well-engineered processing strategies, such as an effective multichannel SE including dereverberation, deep neural network (DNN)-based acoustic modeling, and acoustic model adaptation. While the performance of the challenge baseline GMM-HMM system with multi-condition training and constrained maximum likelihood linear regression (CMLLR) achieved a WER of 49.2 % for the real recordings, the best performing system achieved 9.0 % using eight microphones. The SE task results reveal that we can now effectively reduce the perceived amount of reverberation in both the single-channel and multichannel scenarios and simultaneously improve the overall sound quality, even in severely reverberant real environments. In addition, after analyzing the relationship between the results of the listening and instrumental tests, we show that even though a subjective judgment of the perceived amount of reverberation can be roughly captured with instrumental measures, the overall sound quality cannot be well represented with the metrics employed in this challenge.

1.3 Organization

The remainder of this paper is organized as follows. In Sections 2 and 3, we describe the challenge's design. Section 2 details the problem posed by the challenge and reviews the test datasets. Section 3 introduces its two tasks, SE and ASR, and the evaluation metrics used in each one. In Section 4, we provide an overview of the submitted systems and their key components. Sections 5 and 6 present the results obtained from the ASR and SE tasks. We analyzed the results to identify trends, reveal the state-of-the-art, and clarify the remaining challenges faced by reverberant speech-processing research. The paper is concluded in Section 6.

2 Dataset

The challenge assumes scenarios in which an utterance spoken by a single spatially stationary speaker in a reverberant room is captured with single-channel (1-ch), two-channel (2-ch), or eight-channel (8-ch) circular microphone arrays (Fig. 1). As a part of the challenge, we provided a dataset that consists of a training set, a development (Dev) test set, and an evaluation (Eval) test set, all of which were provided as 1-ch, 2-ch, and 8-ch recordings at a sampling frequency of 16 kHz. All of the data related to the challenge are available through the challenge webpage [7] in its "download" section. Although the specifications of the challenge data have been summarized [6, 7], we briefly review them here for completeness. An overview of all the datasets is given in Fig. 2. Details of each one are given in the following subsections.

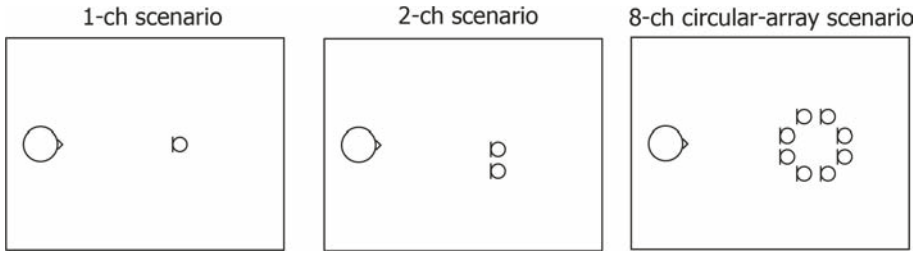

Fig. 1 Target scenarios of REVERB challenge. Here, distances among adjustment microphones are all 7.5 cm

2.1 Test data: development and evaluation sets

When preparing the test data, we took special care regarding the following points:

- The test data have to allow the challenge participants to thoroughly evaluate their algorithms for (i) practicality in realistic conditions and (ii) robustness against a wide range of reverberant conditions. To simultaneously fulfill these conditions, both the Dev and Eval test sets were designed to consist of real recordings (RealData) and simulated data (SimData) (Fig. 2).
- To allow a degree of comparison between SimData and RealData, part of the former was designed to have similar characteristics to the latter in terms of acoustic conditions, i.e., reverberation time and speaker-microphone distance, and complexity of utterance content, i.e., text prompts.

Based on the above design concepts for the test data, Simdata and RealData were prepared as follows:

- **SimData** is comprised of reverberant utterances generated based on the WSJCAM0 corpus [11]. 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:

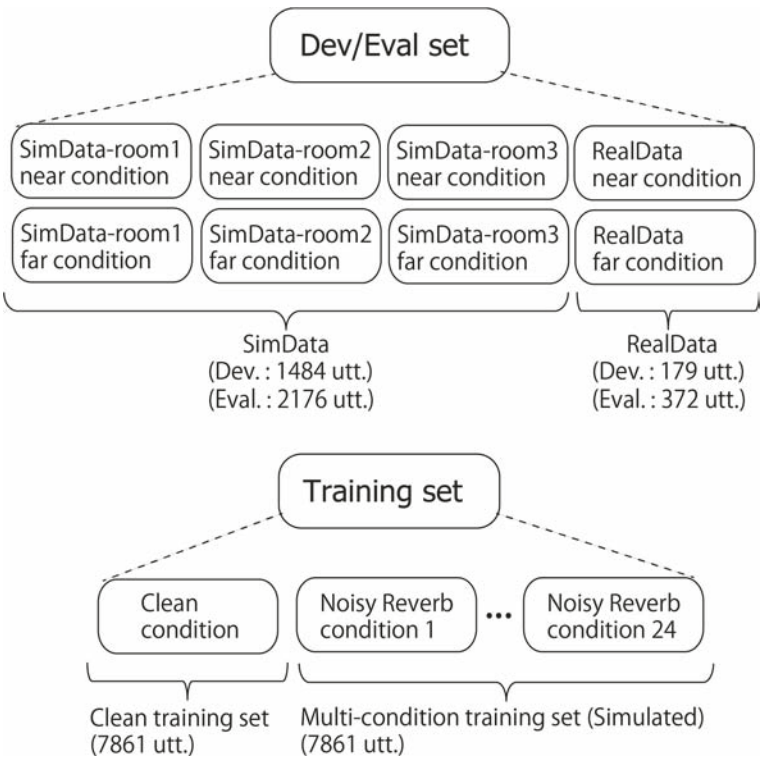

Fig. 2 Overview of datasets used in REVERB challenge. Average durations of utterances in training and test sets are about 7.5 and 6.9 s, respectively

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 are about 0.3, 0.6, and 0.7 s, respectively. The direct-to-reverberation ratios (i.e., D_{50}) for SimData-room1 near and far, -room2 near and far, and -room3 near and far conditions are 99, 98, 95, 79, 97, and 81 %, respectively. D_{50} refers to the percentage of the energy of the direct path plus early reflections up to 50 *ms*, relative to the total energy of the RIR. The RIRs and added noise were recorded in the corresponding reverberant room at the same position with the same microphone array, an 8-ch circular array with a diameter of 20 cm. The array is equipped with omni-directional microphones. 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 is comprised of utterances from the MC-WSJ-AV corpus [12], consists of utterances spoken by human speakers in a noisy and reverberant room. Consequently the sound source cannot be regarded as completely spatially stationary due to the speaker's head movements. The room used for the RealData recording is different from the rooms used for SimData. The room's reverberation time was about 0.7 s [12]. The recordings contain some stationary ambient noise, which was mainly caused by the air conditioning systems. RealData contains two reverberation conditions: one room and two distances between the speaker and the microphone array (near∼100 cm and far∼250 cm). The recordings were measured with an array whose geometry is identical as that used for SimData. Judging by the reverberation time and the distance between the microphone array and the speaker, RealData's characteristics will probably resemble those of the SimData-room-3-far condition. The text prompts of the utterances used in RealData and in part of SimData are the same. Therefore, we can use the same language and acoustic models for both SimData and RealData.

For both SimData and RealData, we assumed that the speakers stay in the same room for each test condition. However, within each condition, the relative speaker-microphone position changes from utterance to utterance. Note that the term "test condition" in this paper refers to one of the eight reverberation conditions that comprise two conditions in RealData and six conditions in SimData (Fig. 2).

2.2 Training set

As shown in Fig. 2, the training dataset consists of (i) a clean training set taken from the original WSJCAM0 training set and (ii) a multi-condition (MC) training set, which 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 measured impulse responses for this dataset range roughly from 0.2 to 0.8 s. Different recording rooms were used for the Dev set, the Eval set, and the training data.

3 Tasks in REVERB challenge

The REVERB challenge consists of two tasks: one for SE and another for ASR, both of which are based on the dataset explained in the previous section. The following subsections describe the details of each task and the evaluation metric(s) employed.

3.1 ASR task

The ASR task is to recognize each test reverberant 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 have to perform recognition without knowing which speaker is talking in which acoustic condition. A baseline ASR system was provided. The baseline system, which is based on HTK, is a triphone GMM-HMM recognizer trained on clean/multi-condition training data. It also includes a function to perform CMLLR-based adaptation. The language model was a bigram scheme. Participants were allowed to take part in either (or both) single-channel and multichannel tasks by employing any input features, acoustic models, training criteria, decoding strategies, and advanced single-channel/multichannel front-end processing technologies, which could be completely different from the challenge baseline ASR systems. Although the relative speaker-microphone position changed randomly from utterance to utterance, the participants were allowed to use all the utterances from a single test condition and to perform full-batch processing. Thus, they could perform, e.g., multiple passes of unsupervised adaptation on the data of a single test condition until the final results are achieved. The world error rate (WER) was used as an evaluation metric.

3.2 SE task

For the SE task, the participants were allowed to participate in either (or both) the single-channel and multichannel tasks using their speech enhancement algorithms. Processed signals were evaluated by listening tests and several different instrumental measures summarized in the following subsections. This evaluation approach was

taken because no universally accepted set of instrumental measures has yet been fully established for evaluating dereverberation algorithms. The SE task is designed not only to reveal the relative merits and demerits of different SE approaches but also to elucidate the characteristics of each instrumental measure, which may facilitate the future research and development of SE algorithms.

3.2.1 Instrumental test

The following instrumental measures were employed: frequency-weighted segmental SNR (FWSegSNR) [13], cepstral distance (CD) [13], log-likelihood ratio (LLR) [13], speech-to-reverberation modulation energy ratio (SRMR) [14], and optionally PESQ [15]. The metrics FWSegSNR, CD, LLR, and PESQ were selected because they correlated well with the listening test results for evaluating the overall quality of the signals processed by various speech enhancement algorithms [13]. The SRMR metric was selected because it concentrates on measuring the dereverberation effect and is non-intrusive unlike the others. This is a favorable characteristic especially when we have no access to reference clean speech signals but only to the observed signals.

3.2.2 Listening test

The audible quality of the processed signals was evaluated in the framework of a multiple stimuli with hidden reference and anchor (MUSHRA) test [16]. Researchers in the speech signal processing field were asked to participate in the test in a crowdsourcing manner. Because of time constraints, we chose this non-standardized listening test style, although it contains the following limitations. For instance, although the subjects were instructed to use headphones in a quiet room, neither the quality of the headphones nor the background noise level in the listening room could be controlled. In addition, it could not be guaranteed that all the subjects had normal hearing.

During the test, all the subjects were first guided to training sessions in which they familiarized themselves with the listening test. Following the training sessions, in each test session, a subject compared a reference sound excerpt (i.e., a clean or headset recording) and a number of test sound excerpts that included an unmarked reference sound (serving as a hidden reference), a noisy reverberant sound (serving as an anchor signal), and processed versions of the same utterance. The following two metrics were used to evaluate the audible quality:

- **Perceived amount of reverberation**: This metric, which represents the perceptual impressions of the degree to which the reference and test sound excerpts are reverberant, assessed the degree of dereverberation a system performed.
- **Overall quality**: This metric evaluated the "sound quality" in a general sense. Subjects gave ratings based on their own judgment regarding any and all detected differences (in terms of naturalness, processing distortion, timbral and reverberation characteristics, additive noise, and so on) between the reference and test sound excerpts.

Figure 3 shows examples of the computer screen used for testing each listening test attribute. The grading scale

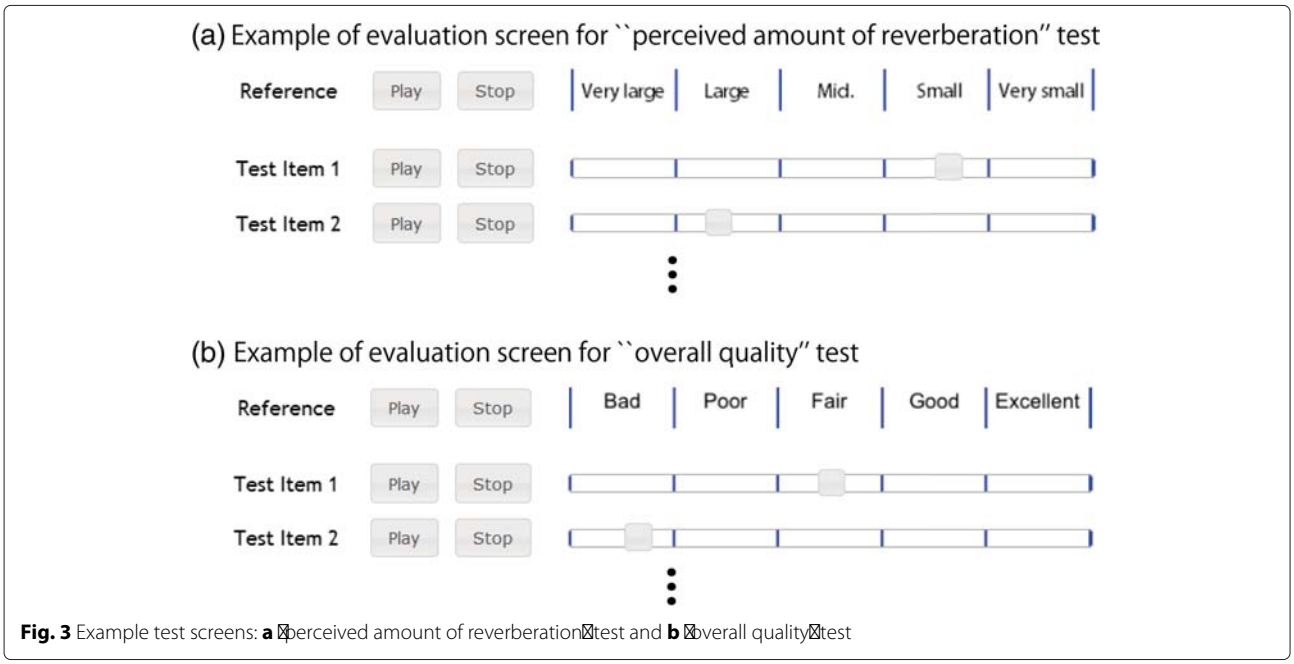

Fig. 3 Example test screens: **a** "perceived amount of reverberation" test and **b** "overall quality" test

ranged from "very large" to "very small" in the "perceived amount of reverberation" test and from "bad" to "excellent" in the "overall quality" test. The subjects used sliders, such as those depicted in Fig. 3, to record their ratings of each test item. As in the standard MUSHRA, prior to the test, the subjects were informed that the reference sound excerpt (i.e., a clean or a headset recording) was hidden among the test items as a hidden reference. They were asked to find it and to give it a high-end rating, i.e., "very small" (or "excellent").

In the test, 1-ch, 2-ch, and 8-ch systems were evaluated separately, since the number of test items was too large to evaluate all of them together. All submitted systems were first regrouped into categories (i.e., 1-ch, 2-ch, and 8-ch) according to the number of microphones they employed. All systems from a given category were assigned to a single test session, meaning that a subject was asked to evaluate all systems from a given category in an assigned test session. The systems were evaluated under four different test conditions: SimData-room2 near and far and RealData near and far. RealData was selected to evaluate the systems in realistic severe reverberation conditions, while SimData-room2 was selected to perform evaluation in moderate reverberation conditions. Evaluations in other conditions, SimData-room1 and -room3, were omitted due to time constraints. For each reverberation condition, two female and two male utterances were randomly selected as test materials. In total, 48 sessions were prepared, i.e., three groups of systems (1-ch, 2-ch, 8-ch), four types of utterances (two females, two males) and four reverberation conditions (RealData near and far, SimData-room2 near and far). Each subject was assigned to one of the 48 sessions and evaluated all systems assigned to the session for the perceived amount of reverberation and overall quality.

4 Submitted systems

Twenty-seven papers were submitted to the REVERB challenge [17–43], which include 25 systems for the SE task and 49 systems for the ASR task. In general, each submitted system had all or a subset of the components shown in Fig. 4. The participants in the SE task mainly focused on the development of the enhancement part in Fig. 4, but the ASR task participants focused on both the enhancement and recognition parts.

Table 1 summarizes information about which task(s) each participant addressed as well as the characteristics of the enhancement and recognition system(s) proposed in each submission. Note that one submission often proposed more than one system. In such cases, if one of the proposed systems in a submission adopted an attribute listed in Table 1, the submission was marked with an "x" under the corresponding feature category.

4.1 Algorithms related to enhancement part

This subsection summarizes the characteristics of the SE/feature enhancement (FE) algorithms submitted to the challenge, which correspond to the components in the enhancement part in Fig. 4. Here, rather than listing all the SE/FE components of the submitted systems, we highlight the methods that effectively dealt with reverberation, based on the challenge results that will be detailed later.

4.1.1 STFT-domain inverse filtering methods

A method proposed in [20] effectively dealt with reverberation by adopting accurate RIR modeling, i.e., convolution in the short-time Fourier transformation (STFT) domain, and removing the distortion by inverse filtering to correct both the amplitude and the phase information. To estimate the inverse filter, a weighted linear prediction error (WPE) method was utilized [20]. The WPE algorithm performs long-term linear prediction at each frequency bin in the STFT domain as follows:

$$\boldsymbol{y}_n[f] = \sum_{\tau=T_\perp}^{T_\top} \boldsymbol{G}_\tau[f]^H \boldsymbol{y}_{n-\tau}[f] + \boldsymbol{e}_n[f], \tag{2}$$

where $\boldsymbol{y}_n$ is a vector comprised of the STFT coefficients of single/multiple microphone signals, $\boldsymbol{e}_n$ is the prediction error vector, $\boldsymbol{G}_\tau$ is a complex-valued square matrix, called prediction matrix, n is the time frame index, $T_\top$ and $T_\perp$ are integers with $T_\top > T_\perp > 0$, and superscript H is a conjugate transposition. Note that, due to the time-varying nature of speech, clean speech signal is not correlated with its past samples (after some delay $T_\perp$). Since late reverberation components are generated from reflections of the past speech samples, they are uncorrelated with the present speech signal. Therefore, linear prediction can only predict the late reverberation and not the clean speech signal component, which will remain as the prediction error/residual. Accordingly, the term $\sum_{\tau=T_\perp}^{T_\top} \boldsymbol{G}_\tau[f]^H \boldsymbol{y}_{n-\tau}[f]$ represents the late reverberant components contained in microphone signals $\boldsymbol{y}_n[f]$, and $\boldsymbol{e}_n[f]$ corresponds to the mixture of clean speech signal and early reflection components. The prediction matrices are optimized for each utterance by minimizing the power of an iteratively re-weighted prediction error. Dereverberated signals $\hat{\boldsymbol{s}}_n$ can be obtained as prediction errors

$$\hat{\boldsymbol{s}}_n[f] = \boldsymbol{y}_n[f] - \sum_{\tau=T_\perp}^{T_\top} \boldsymbol{G}_\tau[f]^H \boldsymbol{y}_{n-\tau}[f]. \tag{3}$$

One attractive characteristic of this approach is that it suppresses only the late reverberation components of the observed signal and virtually shortens the room impulse responses between a speaker and microphones by linear time-invariant inverse filtering, as seen in Eq. (3). Since the algorithm can keep the direct path and early reflection

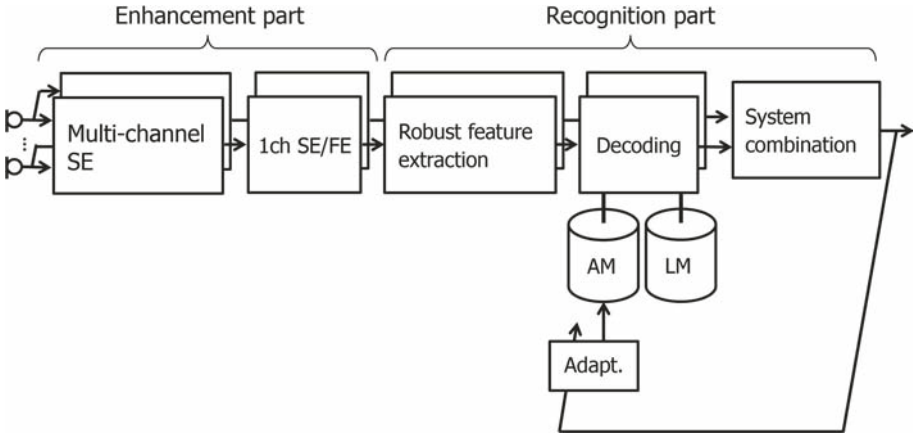

Fig. 4 General block diagram encompassing most systems submitted to REVERB challenge

Table 1 Overview of processing strategies employed by submitted systems

	Task		Characteristics of enhancement part				Characteristics of recognition part			
	SE	ASR	# of channels			Derev.	Advanced feature	NN-based AM	Feat./model adaption	Advanced decoding
			1-ch	2-ch	8-ch					
Alam [17]		x	x			x	x	x		x
Astudillo [18]		x			x	x			x	
Cauchi [19]	x	x	x		x	x				
Delcroix [20]	x	x	x	x	x	x		x	x	x
Epain [21]	x				x	x				
Feng [22]		x	x		x	x	x		x	
Geiger [23]		x			x	x		x	x	x
Gonzalez [24]	x		x			x				
Hirsch [25]		x	x			x				
Kallasjoki [26]		x	x		x	x			x	
Kondo [27]	x		x			x				
Leng [28]	x	x	x		x	x		x	x	
Lopez [29]	x		x			x				
Mimura [30]		x	x			x		x		
Mitra [31]		x	x				x		x	x
Moshirynia [32]	x		x			x				
Ohtani [33]	x		x			x				
Palomaki [34]		x	x			x			x	
Parada [35]		x	x				x		x	
Tachioka [36]		x	x		x	x		x	x	x
Veras [37]	x		x			x				
Wang [38]	x	x		x		x			x	
Weninger [39]		x	x		x	x		x	x	x
Wisdom [40]	x		x	x	x	x				
Xiao [41]	x	x	x	x	x	x		x	x	
Xiong [42]		x	x			x			x	x
Yu [43]	x				x	x				

components of each microphone unchanged, it preserves essential information such as the time difference of arrival (TDOA) and thus subsequently allows multichannel noise reduction techniques based on beamforming to be effectively performed.

4.1.2 Methods based on non-negative RIR modeling

Many submissions utilized a 1-ch algorithm that models the convolutional effect of reverberation in the amplitude domain [17, 26, 29, 32, 43] and showed its efficacy. They assumed that at each frequency bin f, the observed amplitude spectrum $Y_n[f]$ at frame n is generated by the convolution of the amplitude spectra of clean speech $S_{n-M-1}[f], \ldots, S_n[f]$ and those of an RIR $H_0[f], \ldots, H_{M-1}[f]$ as

$$Y_n[f] = \sum_{m=1}^{M-1} S_{n-m}[f]\, H_m[f]\,. \tag{4}$$

Although the potential maximum performance of this type of approach may not be as high as the above inverse filtering approaches due to the non-negative approximation in RIR modeling, such types might be more robust against additive noise and other unexpected distortions because approaches which correct only amplitude information are in general more robust than the ones which aim to correct both the amplitude and phase information. A popular approach in this category is based on non-negative matrix factor deconvolution (NMFD) [26, 32, 43], in which the above equation is expressed using matrix convolution with a shift operator "$m \rightarrow$" as:

$$\boldsymbol{Y} = \sum_{m=0}^{M-1} \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}, \tag{5}$$

$$\boldsymbol{Y} = \begin{pmatrix} Y_1[1] & \cdots & Y_N[1] \\ \vdots & \ddots & \vdots \\ Y_1[F] & \cdots & Y_N[F] \end{pmatrix}, \tag{6}$$

$$\boldsymbol{H}_m = \begin{pmatrix} H_m[1] & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & H_m[F] \end{pmatrix}, \tag{7}$$

$$\boldsymbol{S} = \begin{pmatrix} S_1[1] & \cdots & S_N[1] \\ \vdots & \ddots & \vdots \\ S_1[F] & \cdots & S_N[F] \end{pmatrix}, \tag{8}$$

where F and N correspond to the total number of frequency bins and the total number of observed frames. The shift operator "$m \rightarrow$" shifts the columns of its argument by m positions to the right:

$$\overset{0\rightarrow}{\boldsymbol{S}} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}, \quad \overset{1\rightarrow}{\boldsymbol{S}} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{pmatrix}. \tag{9}$$

Entries of the matrics $\boldsymbol{Y}$, $\boldsymbol{H}_m$, $\boldsymbol{S}$ are all non-negative. The parameter M is chosen to be sufficiently large such that it can cover the energy of reverberation. NMFD decomposes the observed amplitude spectrogram $\boldsymbol{Y}$ into the convolution of the amplitude domain RIR $\boldsymbol{H}_0, \ldots, \boldsymbol{H}_{M-1}$ and the clean speech spectrogram $\boldsymbol{S}$. Some research [26, 32] further decomposed estimated clean spectrogram $\boldsymbol{S}$ based on the non-negative matrix factorization (NMF) concept. By doing so, they introduced a widely used NMF-based speech model, i.e., a pretrained dictionary of the clean amplitude spectrum, to the NMFD-based dereverberation framework, which allows them to perform semi-supervised speech enhancement. With such a pretrained dictionary, the clean speech characteristics in dereverberated signals can be preserved.

4.1.3 Methods based on statistical RIR modeling

Another widely used effective 1-ch approach employed a simple statistical model for the RIRs [44]. In this approach, the RIR $h(t)$ is modeled as white noise modulated by an exponentially decaying envelope whose decay rate is determined by the reverberation time [44] as follows:

$$\begin{aligned} h(t) &= \begin{cases} a(t)e^{-\Delta t}, & \text{for } t > 0 \\ 0, & \text{(otherwise)} \end{cases} \\ \Delta &= \frac{3\ln(10)}{\mathrm{RT}_{60}}, \end{aligned} \tag{10}$$

where $a(t)$ is a zero-mean white noise sequence with variance σ_a^2 and RT_{60} is the reverberation time.

Assuming that the observation is generated through the time-domain convolution of clean speech with this simplified RIR, an estimate of the reverberation's power spectrum at the n-th frame, $|R_n[f]\,|^2$, is obtained simply by weighting the observed power spectrum at past frame $|Y_{n-K}[f]\,|^2$ as

$$|\hat{R}_n[f]\,|^2 = e^{-2\Delta T_d} |Y_{n-K}[f]\,|^2. \tag{11}$$

Here, $K = \lfloor T_d f_\mathrm{s}/\lambda \rfloor$, and T_d is generally set roughly to 50 ms. λ denotes the frame shift of the STFT in samples. Dereverberated speech is then obtained by subtracting the estimated reverberant power spectrum $|\hat{R}_n[f]\,|^2$ from the observed power spectrum $|Y_n[f]\,|^2$ as in spectral subtraction [19, 36–38, 41, 42]. Alternatively, some extensions have also been proposed to this approach, e.g., analysis and synthesis in the short-time fan-chirp transform domain [40]. The apparent advantages of this approach are its low computational complexity and robustness against noise.

4.1.4 Methods based on nonlinear mapping

Some submissions used an approach in which no explicit reverberation model was assumed. In this type of approach, stereo training data are used to learn a nonlinear mapping function between noisy reverberant and clean speech. Typical approaches in this category include

a denoising auto-encoder (DAE) that uses fully connected feed-forward DNNs [30, 41] or bidirectional long short-term memory (BLSTM) recurrent neural networks (RNNs) [39]. Given a sequence of the input features (e.g., log Mel-filterbank features), $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$, it estimates the output feature at the n-th frame $\hat{\mathcal{S}}_n$ based on a pretrained highly nonlinear mapping function, i.e., a neural network, as

$$\hat{\mathcal{S}}_n = \mathcal{F}\{\{\mathcal{Y}_1, \ldots, \mathcal{Y}_N\}; \theta\}, \tag{12}$$

where $\mathcal{F}\{\cdot; \theta\}$ represents a nonlinear transformation based on a neural network with parameters θ. Although such an approach is guaranteed to work effectively if the test and training conditions are matched, it is very interesting to determine whether it can be generalized to unseen acoustic conditions. The challenge results indicate that DAE can be generalized to handle RealData which is quite different from the DAE training data. An advantage of these approaches is that since they work in the same feature domain as ASR systems, they can be smoothly integrated with any back-end system. It is also possible to integrate them tightly with DNN-based acoustic models by optimizing θ jointly with the acoustic model parameters based on the same constraint as the ASR systems.

4.2 Algorithms related to the recognition part

This subsection summarizes the characteristics of the ASR algorithms submitted to the challenge that correspond to the components in the recognition part in Fig. 4. In Table 1, the recognition part of each submission is characterized with respect to the presence/absence of the following strategies:

- Advanced features (e.g., i-vector [17], gammatone cepstral coefficient [17])
- Deep neural network (DNN)-based acoustic model (AM) [17, 20, 23, 28, 30, 36, 39, 41]
- Feature/model-space adaptation (e.g., maximum likelihood linear regression (MLLR) [18, 22, 23, 26, 28, 29, 31, 34, 36, 38, 41, 42], modified imputation [18], layer adaptation of DNN [20])
- Advanced decoding (e.g., recognizer output voting error reduction (ROVER) [17, 18, 31, 36, 42], minimum Bayes risk decoding [36], recurrent neural network-based language model [20])

In general, the top-performing systems employed quite advanced techniques regarding these processing strategies. However, all the recognition approaches employed in the challenge, i.e., robust features, acoustic modeling scheme, feature/model-space adaptation, advanced decoding, are not the ones designed specifically for reverberation robustness, but rather for general robustness purposes. Therefore, for conciseness, we omit a detailed description of the ASR techniques employed in the submitted systems. However, note that the challenge results, which will be detailed later, indicate that higher ASR performance was achieved not only with a powerful enhancement algorithm but also with the advanced ASR strategies for acoustic modeling, adaptation, and decoding techniques.

5 ASR results and related discussions

In this section, we present the ASR results of all 49 systems submitted to the challenge. Then, we scrutinize the data to uncover hidden trends and provide insights about effective processing strategies for the reverberant speech recognition task. Finally, building on the findings of this analysis, we summarize current achievements and the remaining challenges of this task.

5.1 Overall results

The overall results of the ASR task are presented in Fig. 5. To make the comparisons as fair as possible, we grouped the submitted results by the processing conditions (i.e., number of microphones, data used for acoustic model training) employed in each system and presented them in one of nine panels in Fig. 5. The vertical axes have a logarithmic scale for the sake of visibility. Panels (a) to (c) show the results obtained based on acoustic models trained with the clean training data. Panels (a), (b), and (c) correspond to the results based on 1-ch, 2-ch, and 8-ch processing schemes, respectively. The results presented in panels (d) to (f) correspond to the results obtained with acoustic models trained with the multi-condition data provided by the challenge. The results presented in panels (g) to (i) were obtained with acoustic models trained with the extended (multi-condition) data prepared by each participant. Interactive graphs of the overall results can be found on the challenge webpage [7]. As mentioned above, one submission often proposed more than one system and submitted multiple results to the challenge. To handle these cases in a simple manner in Fig. 5, such multiple results belonging to one submission are indicated with the same colored line under the name of each submission. We summarized the overall trends of the results in Appendix A. In the following, we focus more on the top-performing systems [20, 22, 36, 39] to determine the essential components to achieve the lowest WERs.

5.2 Key components in systems achieving lowest WERs

In this subsection, we focus on the analysis of the systems that achieved the lowest WERs by average RealData scores [20, 22, 36, 39]. We first discuss the ideas shared by these

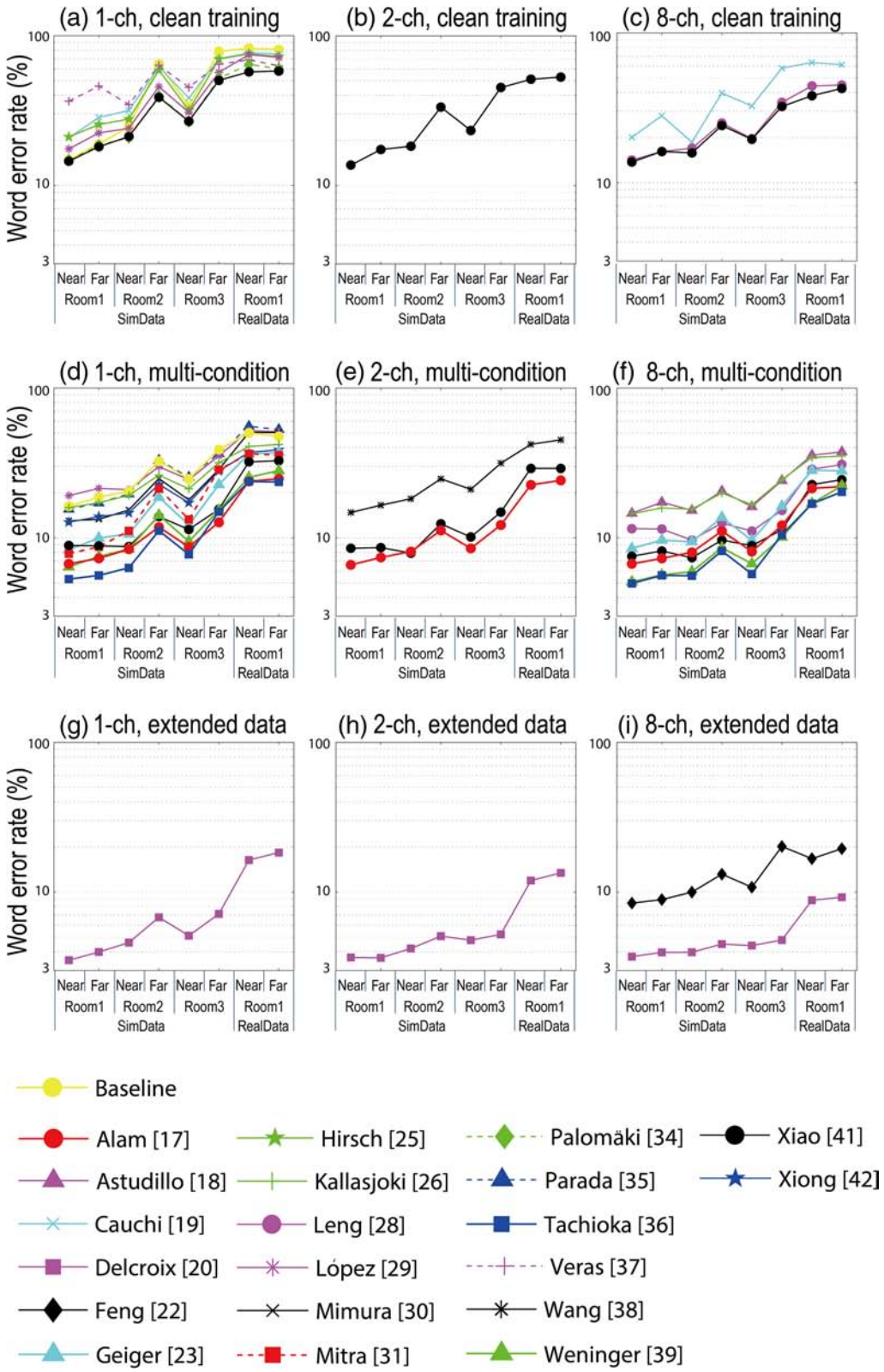

Fig. 5 WERs of submitted systems, listed separately by number of microphones and data used for acoustic model training

systems and then briefly review the key components used in each of the top-performing systems.

Most top-performing systems employed advanced technique(s) in all or some of the following processing components, each of which contributed to significantly reduce WER.

- Speech/feature enhancement such as beamforming and inverse filtering
- Advanced acoustic modeling such as DNN
- Acoustic model adaptation

More specifically, in [20, 22, 36, 39], they commonly focused on employing (1) beamforming and/or dereverberation techniques that utilize multichannel acoustic diversity, (2) powerful acoustic models such as DNN or a subspace Gaussian mixture model (SGMM) trained with discriminative training criteria, and (3) acoustic model adaptation techniques to mitigate the mismatch between the training data and the signal processed by the SE/FE front-end. The fact that the above processing architecture and ideas are common to all the top-performing systems might implicitly indicate that these key components should be jointly utilized and optimized to achieve the lowest WERs.

This finding certainly coincides well with previous studies. For example, it was already shown that beamforming techniques can greatly improve the performance of distant speech recognition even when used with powerful DNN-based acoustic models [45]. An interesting finding, which may be unique to the REVERB challenge results, is that since the top-performing systems employed dereverberation techniques; dereverberation in addition to beamforming is necessary to achieve high recognition performance in severe reverberant and noisy environments. Moreover, although it was already known that DNN-based acoustic models outperform legacy GMM-HMM models under environments with additive noise and channel distortions [45, 46], they also work well in reverberant environments. Various acoustic model adaptation schemes were also found effective when jointly used with front-end processing including dereverberation.

Next, we briefly describe the characteristics of each top-performing system [20, 22, 36, 39] and reveal to the highest extent possible why in particular these systems worked well.

- The front-end processing of the system proposed by Delcroix et al. [20] employed linear prediction-based multichannel dereverberation (introduced in Section 4.1.), followed by minimum variance distortionless response (MVDR) beamforming. The use of the multichannel dereverberation technique allows them to exploit multi-microphone acoustic diversity for both dereverberation and beamforming. Moreover, filtering operation of these front-end processings are completely linear so that they did not introduce unfavorable nonlinear distortion to the processed signal. Their result shows that the 8-ch dereverberation achieved more than 30 % relative WER reduction (RWERR), while MVDR beamforming also achieved about 30 % RWERR when they are used with a DNN-based acoustic model. In their back-end, they showed that just by changing the baseline GMM-HMM acoustic model to DNN and introducing a trigram language model, they achieved about 60 % RWERR. In addition, adapting a layer of the DNN model brought about 15 % RWERR.
- Tachioka et al. [36] employed simple but robust front-end processing for steady improvement and focused more on strong acoustic models that were combined with various advanced training and adaptation schemes. In their multichannel front-end system, they first applied delay-sum beamforming to the input signal before the statistical RIR-based 1-ch dereverberation technique introduced in Section 4.1.3. The delay-sum beamformer achieved about 10 % RWERR, and dereverberation achieved a few percent of RWERR. In their systems, adaptation schemes such as feature-space MLLR and maximum likelihood linear transformation (MLLT) greatly contributed to the improvement and achieved about 30 % RWERR. They used a unique technique called a dual system combination to construct various (> 10) complementary acoustic models and combined their outputs using ROVER, which contributed to about 7 % RWERR.
- Weninger et al. [39] employed a feature enhancement scheme based on a state-of-the-art neural network, i.e., BLSTM-based DAE introduced in Section 4.4, and achieved good performance, combining it with back-end systems that employ a number of feature transformation and adaptation techniques. Their front-end system achieved substantial improvement, i.e., more than 30 % RWERR, when a BLSTM-based DAE was combined with a simple 8-ch delay-and-sum beamformer. This improvement was obtained based on a strong GMM-HMM back-end system combined with feature adaptation techniques such as feature-space MLLR and BMMI-based discriminative training.
- Feng et al. [22] strongly focused on multiple passes of unsupervised feature- and model-space speaker adaptation using CMLLR, MLLR, and vocal tract length normalization (VTLN). Combining such techniques with their front-end beamformer, they achieved a total of more than 70 % RWERR.

Note that a common characteristic of the systems that achieved the lowest WERs is that their performance was achieved not as the result of a single standout algorithm but through the careful combination of multichannel front-end processing, strong acoustic modeling, and feature-/model-space adaptation.

5.3 Current achievements and remaining challenges for ASR

Figure 6 shows the WERs obtained by the 1-ch, 2-ch, and 8-ch top-performing systems. The vertical axis corresponds to the averaged WERs of RealData and the horizontal axis to those of SimData. The striped rectangular area indicates recognition errors that might not be related to the environmental distortions. This region is determined by the recognition rate of the clean/headset speech (SimData; 3.5 %, RealData; 6.9 %), which was obtained with a state-of-the-art DNN-HMM speech recognizer [20].

From Fig. 6 and the previous section, we can summarize the current achievements and the remaining challenges as follows:

- Although the multichannel systems, especially the 8-ch systems, closely approached the clean/headset performance, the 1-ch systems remain greatly inferior, suggesting considerable room for future improvement. Since the 8-ch algorithms generally impose severe hardware constraints on the overall system and are impractical in many situations, 1-ch algorithms must achieve the recognition performance currently achieved by 8-ch systems.

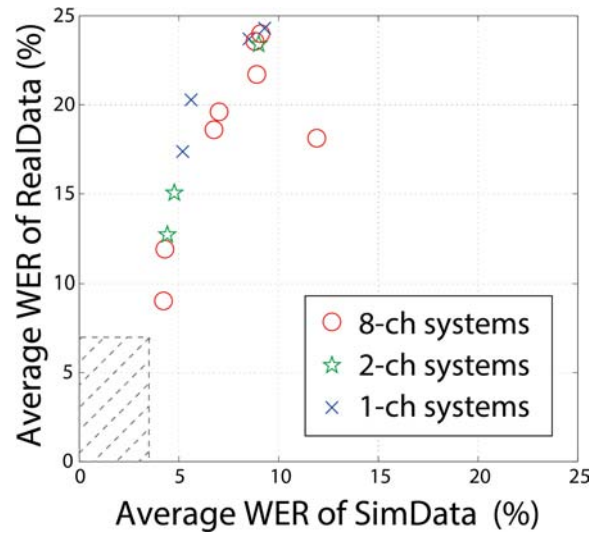

Fig. 6 Scatter plot showing WERs of top-performing 1-ch/2-ch/8-ch systems. The *Striped rectangular area* indicates errors caused by factors other than environmental distortions

- The top-performing systems introduced in the previous section accomplish their current level of performance by repeatedly processing the input data by several enhancement algorithms and performing multiple passes of feature-/model-space adaptation. However, since many ASR applications require real-time/online processing, pursuing research on such processing schemes is critical.
- Apart from the problems of ASR techniques, concerning the challenge data preparation stage, challenges remain in simulating acoustic data that are close to actual recordings. The results obtained with SimData-room3 and RealData are strongly correlated on a gross level, as shown in Appendix B. But, Figs. 5 and 6 show that although the acoustic conditions simulated with SimData-room3 are supposed to be close to RealData, their WER performances are very different if we only look at the top-performing systems [20, 22, 36, 39]. 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.

6 SE results and related discussions

In this section, we first present the overall SE results in terms of instrumental measures and then briefly mention the single- and multichannel algorithms that achieved good scores and their relations to the ASR results. Finally, we present the results of a listening test and discuss their general tendencies.

6.1 Instrumental test results

In this subsection, we describe the instrumental test results of the SE task. Because of space limitations, we only present the results in terms of FWSegSNR, which represent the general tendencies well that were observed in the instrumental test and coincides well with the listening test results in terms of the perceived amount of reverberation. Please refer to the challenge's webpage [7] for the complete results.

Figure 7 separately summarizes the FWSegSNR results of the 1-ch, 2-ch, and 8-ch systems. In general, it most successfully improved their performance. Some 1-ch systems had difficulty with SimData-room1 where the reverberation level was quite low. Not surprisingly, the systems that employed multichannel processing tended to perform better in most conditions.

Next, we briefly mention the single- and multichannel algorithms that achieved high FWSegSNRs, describing how they are different/similar to the other systems and their relations to the results obtained from the ASR task. The following single-channel algorithms achieved high FWSegSNRs [32, 41]:

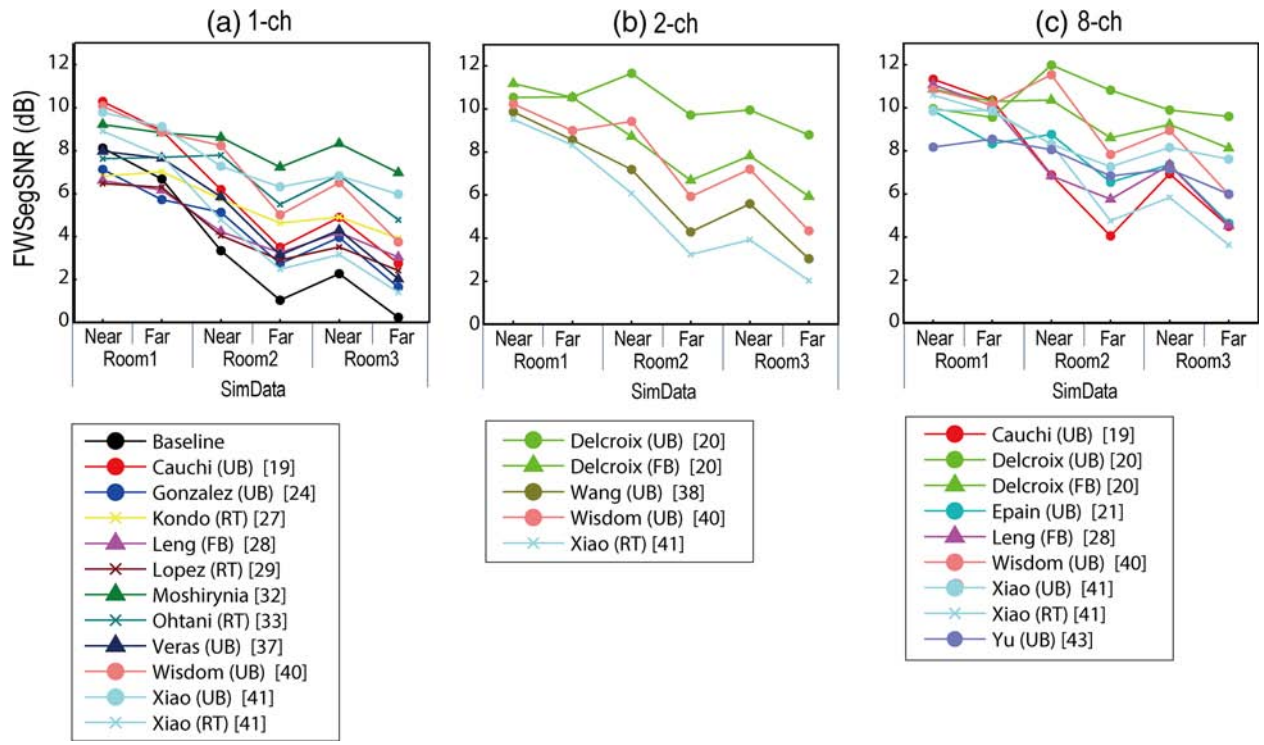

Fig. 7 Frequency-weighted Segmental SNR of submitted systems, listed separately by number of microphones

- Moshirynia et al. [32] showed that NMFD combined with joint dictionary learning, introduced in Section 4.1.2, works well for 1-ch dereverberation. In their algorithm, they first applied NMFD to remove the late reverberation components, which were caused by the signals from the past frames, and subsequent joint-dictionary-based NMF removed the early reflection effect. The joint dictionary used for NMF learned pairs of exemplars with and without the early reflection effect, and thus it can map a signal that contains the early reflection effect, i.e., the signal processed by NMFD, to the one without the early reflection effect. Note that this technique is a relatively rare method that can remove both late and early reflections.
- Xiao et al. [41] employed statistical RIR-based dereverberation (Section 4.1.3). Interestingly, although some submissions [19, 40] employed the same or similar methods, they achieved lower FWSegSNR scores, possibly due to implementation issues or parameter tuning strategies.

Similarly, the following multichannel systems achieved high FWSegSNRs [20, 40]

- Delcroix et al. [20] employed a linear time-invariant inverse filtering method (Section 4.1.1) followed by an MVDR beamformer, which was also found effective for ASR.
- Wisdom et al. [40] proposed a method consisting of beamforming followed by statistical RIR-based 1-ch dereverberation (Section 4.1.3).
  This simple combination was also investigated for the ASR task in a different submission [36] and provided steady improvement.

Great similarity can be found among the methods effective for ASR and the instrumental test.

6.2 Results of listening test and general tendencies

To investigate the relationship between the SE instrumental test results and the actual audible quality, we conducted the listening test described in Section 3.2.2. Figure 8 shows the listening test results of each submitted system. They are based on 126 valid responses for the "perceived amount of reverberation" test and 128 valid responses for the "overall quality" test[1]. We obtained these responses after a post-screening that rejected the responses from subjects who failed to find the hidden reference signal and rated it with a score of less than 95. All the mean scores were plotted with their associated 95 % confidence intervals.

The scores in Fig. 8 are MUSHRA differential scores [47], which are calculated based on the raw MUSHRA

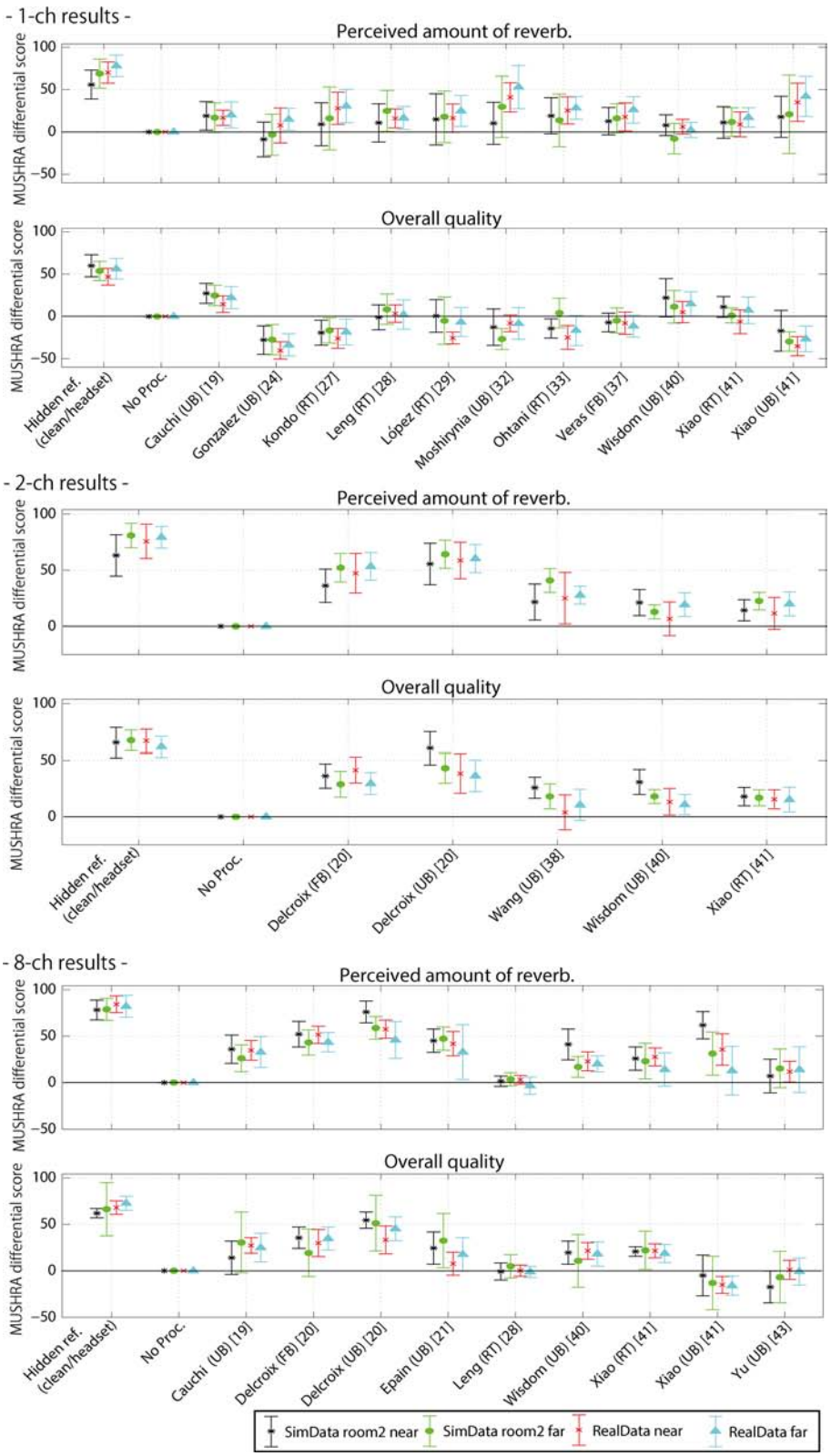

Fig. 8 Listening test results. MUSHRA differential scores for submitted systems under all four test conditions: SimData room-2 near and far and RealData near and far. The *top two panels* show results for all 1-ch systems in terms of the perceived amount of reverberation (*upper panel*) and overall quality (*lower panel*). Two panels in *middle* and *bottom* show results of 2-ch and 8-ch systems

scores obtained from the subjects. As is commonly known, raw MUSHRA scores tend to be significantly biased due to the sense of each subject. For instance, even if two subjects hear exactly the same reverberant sound and are asked about their perceived amount of reverberation, their responses will probably be different. Consequently, a simple average of the raw data without taking these biases into account might result in very large variances, further complicating statistical analysis and decreasing its reliability. To remove such potential biases, we calculated the MUSHRA differential scores by subtracting the scores for the unprocessed signal (hidden anchor signal) from all the other scores.

The top two panels in Fig. 8 show the results for all the 1-ch systems in terms of the perceived amount of reverberation (upper panel) and overall quality (lower panel). The two middle panels show the results for the 2-ch systems, and the bottom two show them for the 8-ch systems. *Directly comparing the numbers among the 1-ch, 2-ch and 8-ch systems should be avoided, since MUSHRA tests were carried out separately for each group of systems.* The scores for each system are composed of four error bars, each of which shows the result obtained in a certain room and under a certain microphone-speaker distance condition. The scores of each system are plotted in conjunction with those of the clean/headset signal (i.e., far left in each panel) and the unprocessed noisy reverberant signal (i.e., indicated as "No proc."). According to the nature of MUSHRA differential scores, the "No proc." scores remained exactly at zero. Thus, if a system has a score significantly higher than zero, its output can be rated significantly better than "No proc.", meaning a lower perceived amount of reverberation or better overall quality.

The listening test results have to be interpreted with great caution, since this test was conducted in a non-standardized crowdsourcing manner, where test conditions such as listening environment and subject quality were not perfectly controlled. With this caution in mind, we conclude that the figure indicates the following tendencies:

- **1-ch systems:** Many systems significantly reduced the perceived amount of reverberation. However, improving the overall quality is more challenging. Among the 1-ch systems, only the one proposed by Cauchi et al. [19] performed significantly better than "No proc." for both metrics.
- **Multichannel systems:** Many systems significantly reduced the perceived amount of reverberation and significantly improved the overall quality. The trends are similar for the 2-ch and 8-ch cases. One of the biggest advantages of multichannel systems is their capability of incorporating linear spatial filtering, which does not induce unnatural nonlinear distortion that might reduce the overall quality.
- **Results under different test conditions:** On the whole, we identified similar trends among the four different test conditions. We found no significant differences among the rankings under each test condition.

6.3 Relationship between listening and instrumental test results

Next, we discuss the validity of the instrumental measures by comparing their scores with the listening test results. Table 2 shows the correlation coefficients that indicate the relationship between the instrumental and listening test results in terms of the perceived amount of reverberation. Table 3 shows the correlation coefficients with the "overall quality" test. We calculated the correlation coefficients separately for each system category (1-ch, 2-ch, and 8-ch). Numbers in the table were obtained by calculating correlation between MUSHRA scores of each system (averaged over all subjects and all sentences) for SimData room-2 near and far conditions and corresponding instrumental test scores.

CD and LLR indicate lower values when the quality is good, unlike the MUSHRA scores. In such cases, strong *negative* correlation indicates that the metrics work appropriately as indicators of audible quality[2].

Table 2 shows the relationship between the instrumental test results and the "perceived amount of reverberation" test. If we compare the rows for the 1-ch, 2-ch, and 8-ch systems, we see that they have similar and consistent values, although there are some minor variations. On average, metrics such as CD and FWSegSNR exhibit a relatively strong correlation and seem to roughly capture the subjectivity regarding the perceived amount of reverberation.

Table 3 shows the relationship between the instrumental test results and the "overall quality" test. In this case, comparing the rows for the 1-ch, 2-ch, and 8-ch systems, we surprisingly find that they take different signs in the 1-ch and multichannel cases. Although for the multichannel systems (especially 2-ch systems), the instrumental measures more or less coincide with the listening test results, the results obtained with the 1-ch systems showed

Table 2 Correlation coefficients between results of instrumental and listening tests in terms of perceived amount of reverberation

	CD	FWSegSNR	LLR	SRMR	(PESQ)
1-ch system	−0.43	0.51	−0.14	0.48	(0.65)
2-ch system	−0.91	0.87	−0.72	0.81	(0.83)
8-ch system	−0.76	0.74	−0.42	0.59	(0.84)

Table 3 Correlation coefficients between results of instrumental and listening tests in terms of overall quality

	CD	FWSegSNR	LLR	SRMR	(PESQ)
1-ch system	0.29	−0.29	0.47	−0.45	(−0.75)
2-ch system	−0.97	0.96	−0.72	0.76	(0.91)
8-ch system	−0.38	0.49	−0.39	0.06	(0.67)

different trends. These results might simply suggest that no instrumental measure adequately captured the subjective sense of the overall quality, especially in the 1-ch systems. Such a claim can be supported, for example, by the fact that the 1-ch system developed by Cauchi et al. [19] is the only one among 11 submitted 1-ch systems that significantly reduced the perceived amount of reverberation and improved overall quality, whereas their system ranked 5th in terms of FWSegSNR score. Listening tests conducted with more subjects and more controlled listening conditions must be carried out in the future to confirm this trend. As a consequence, with the current results, we could not find instrumental measures that represent the subjective sense regarding overall quality.

7 Conclusions

Reverberation is an inevitable problem when a speech signal is captured unobtrusively with distant microphones because it degrades the audible quality of speech and the performance of ASR systems. This paper outlined the achievements of the REVERB challenge, a community-wide campaign that evaluated speech enhancement and recognition technologies in reverberant environments. The REVERB challenge is comprised of two tasks, SE and ASR, both of which are based on the same data including real recordings.

An analysis of the results obtained in the ASR task indicated that the top-performing systems [20, 22, 36, 39] performed better not due to the standout effect of one particular algorithm but rather by carefully combining several powerful processing strategies. More specifically, their processing strategies seem to commonly emphasize the joint utilization of the following:

- front-end processing such as beamforming and dereverberation that effectively utilize multichannel acoustic diversity by linear filtering,
- strong acoustic models such as DNNs, and
- appropriate acoustic model adaptation schemes that mitigate the mismatch between the front- and back-ends.

No single panacea-like algorithm exists that can alone solve the problem of reverberant speech recognition.

Based on the SE task results, we found the following:

- Almost all the systems improved the results of the instrumental measures.
- Based on the listening test results, many 1-ch SE systems still have difficulty improving the overall speech quality in a consistent and significant manner, even though they did manage to reduce the perceived amount of reverberation. However, one well-engineered and carefully tuned enhancement system [19] effected significant improvement in both metrics.
- Many multichannel systems succeeded in significantly reducing the perceived amount of reverberation as well as significantly improving the overall quality.
- Based on an analysis of the relationship between the listening and instrumental test results, although the subjective sense of the perceived amount of reverberation was roughly captured with some instrumental measures, the overall quality could not be represented with any of the instrumental measures tested here. However, larger scale listening tests must be performed to clarify this issue.

Finally, although the development of an algorithm that can reduce the detrimental effect of reverberation is considered one of the toughest remaining challenges in this research field, the REVERB challenge confirmed that significant progress has recently been made and has identified a number of effective and practical solutions. We are confident that the challenge's data and achievements will fuel future research on reverberant speech enhancement and recognition.

Appendix A: General tendencies observed in ASR results

Since the massive number of results presented in Fig. 5 makes it very difficult to extract trends, we converted it into a bubble chart (Fig. 9) to analyze the data from different perspectives. The bubble chart's purpose is to discover what processing schemes significantly impacted the final results on a gross level. Figure 9 shows only the *gross* effect of each processing scheme on the collection of results, which quite often does not reflect the effectiveness of a particular algorithm proposed in a certain submission. Since the validity of each particular algorithm is confirmed experimentally in each submission, we refer to the corresponding papers for a closer look at the effects of the algorithms and schemes.

In Fig. 9, the area of each circle is proportional to the number of systems that fall into the ± 2 % range of WER corresponding to the middle of the circle. The vertical axis shows the average WER of RealData, and the horizontal axis shows the processing conditions. Here, we focused only on the RealData results, since the RealData and SimData results are closely correlated (Appendix B). Eight

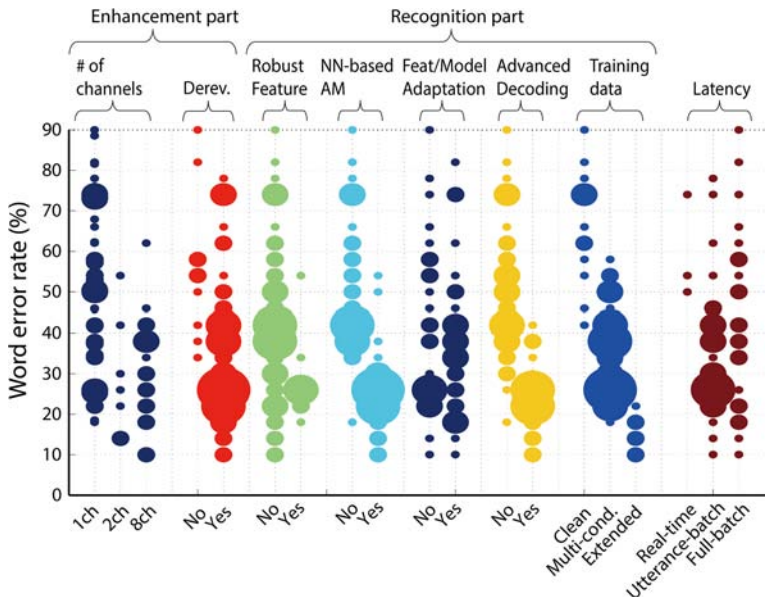

Fig. 9 Bubble chart showing relationship among average WERs of RealData and each processing scheme

bubble charts are shown in the figure, each of which shows the relationship between a WER and the number of microphones (i.e., 1-ch, 2-ch, and 8-ch), the presence/absence of a dereverberation scheme, the presence/absence of robust features, the presence/absence of an NN-based AM, the presence/absence of feature/model-space adaptation, the presence/absence of advanced decoding, the type of training data (i.e., clean, multi-condition, or extended data) and the latency of the proposed system (i.e., real-time (RT), utterance-batch (UB), and full-batch (FB)). The figure indicates the following tendencies:

- The overall results seem greatly influenced by the following two parameters: the type of training data and the presence/absence of DNN-based AM. In the charts that correspond to these two parameters, there is only a slight overlap between the results obtained by the systems that employed these processing schemes and the results obtained by the systems that did not.
- When we focus on the systems that achieved lower WERs, we can see vague trends that show the utility of multichannel processing and the advantage of employing some kind of a dereverberation method and advanced decoding.
- On a gross level, clearly detecting any significant influences of the other parameters is difficult, although each method that corresponds to these parameters was found to be effective in each submission.

Appendix B: Relationship between the SimData and RealData results in the ASR task

When we compared the SimData and RealData results, the systems performed differently with simulated data (i.e., conditions with completely time-invariant RIRs) and real recordings. Figure 10 shows a scatter plot of the results for all the systems, where the vertical and horizontal axes show the WERs of the RealData and SimData-room3 far conditions. The strong positive correlation between the RealData and SimData results indicate that almost all the systems proposed for the REVERB challenge appear to behave similarly for RealData and SimData.

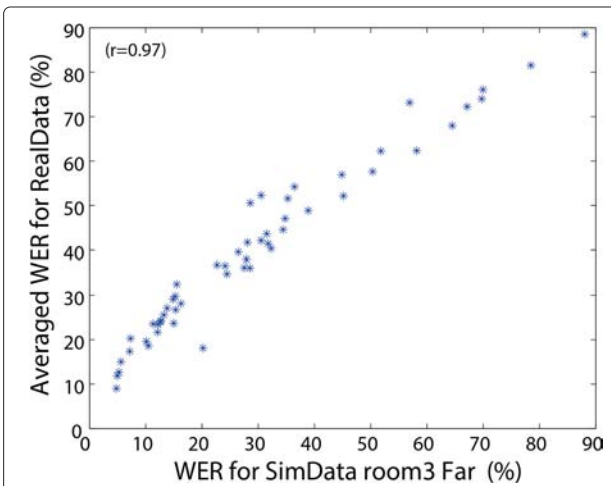

Fig. 10 Scatter plot showing relationship between SimData and RealData results

Endnotes

[1]We used about 30 responses to calculate the average for each condition.

[2]The values related to PESQ are in parentheses; since PESQ was treated as an optional metric in the challenge, we did not collect enough data for it.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors thank the IEEE AASP Technical Committee for supporting this challenge, Christian Hofmann (University of Erlangen-Nuremberg) for his help in measuring room impulse responses, Dr. Tiago Falk (INRS) for providing his SRMR code, Dr. Erich Zwyssig, Dr. Mike Lincoln, and Prof. Steve Renals (University of Edinburgh) for letting us use the MC-WSJ-AV data and their effort to make it available through the LDC. LDC for letting us use the MC-WSJ-AV [48] and WSJCAM0 [49] free of charge within the framework of the evaluation license. The author would also like to thank Google, Audience, TAF, Mitsubishi Electric, DREAMS project, and Nuance for supporting the workshop at which the challenge results were originally disseminated. We also thank the anonymous reviewers for their constructive comments.

Author details
[1]NTT Communication Science Laboratories, Kyoto, Japan. [2]Bar-Ilan University, Ramat Gan, Israel. [3]International Audio Laboratories Erlangen, Erlangen, Germany. [4]University of Paderborn, Paderborn, Germany. [5]Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen-Nuremberg, Germany. [6]Amazon Development Center Germany GmbH, Aachen, (Germany All work done while affiliated with University of Paderborn, Paderborn, Germany). [7]Carnegie Mellon University, Pittsburgh, USA. [8]Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany.

Received: 20 February 2015 Accepted: 7 January 2016

References

1. I Tashev, *Sound capture and processing*. (Wiley, New Jersey, 2009)
2. X Huang, A Acero, H-W Hong, *Spoken language processing: a guide to theory, algorithm and system development*. (Prentice Hall, New Jersey, 2001)
3. M Wölfel, J McDonough, *Distant speech recognition*. (Wiley, New Jersey, 2009)
4. PA Naylor, ND Gaubitch, *Speech dereverberation*. (Springer, Berlin, 2010)
5. T Yoshioka, A Sehr, M Delcroix, K Kinoshita, R Maas, T Nakatani, W Kellermann, Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. **29**(6), 114–126 (2012)
6. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, E Habets, R Haeb-Umbach, V Leutnant, A Sehr, W Kellermann, R Maas, BR S. Gannot, in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech, (2013)
7. REVERB Challenge Webpage. http://reverb2014.dereverberation.com/. Accessed 13 Jan 2016
8. D Pearce, H-G Hirsch, in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, (2000), pp. 29–32
9. E Vincent, S Araki, FJ Theis, G Nolte, P Bofill, H Sawada, A Ozerov, BV Gowreesunker, D Lutter, The signal separation evaluation campaign (2007–2010): achievements and remaining challenges. Signal Process. **92**, 1928–1936 (2012)
10. J Barker, E Vincent, N Ma, C Christensen, P Green, The PASCAL CHiME speech separation and recognition challenge. Comput. Speech and Lang. **27**(3), 621–633 (2013)
11. T Robinson, J Fransen, D Pye, J Foote, S Renals, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition, (1995), pp. 81–84
12. M Lincoln, I McCowan, J Vepa, HK Maganti, in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments, (2005), pp. 357–362
13. Y Hu, PC Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. **16**(1), 229–238 (2008)
14. TH Falk, C Zheng, W-Y Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio, Speech, Lang. Process. **18**(7), 1766–1774 (2010)
15. ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs
16. ITU-R Recommendation BS.1534: method for the subjective assessment of intermediate quality levels of coding systems
17. MJ Alam, V Gupta, P Kenny, P Dumouchel, in *Proceedings of REVERB Challenge Workshop, p2.10*. Use of multiple front-ends and I-vector-based speaker adaptation for robust speech recognition, (2014)
18. RF Astudillo, S Braun, E Habets, in *Proceedings of REVERB Challenge Workshop, o1.3*. A multichannel feature compensation approach for robust ASR in noisy and reverberant environments, (2014)
19. B Cauchi, I Kodrasi, R Rehr, S Gerlach, A Jukić, T Gerkmann, S Doclo, S Goetze, in *Proceedings of REVERB Challenge Workshop, o1.2*. Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme, (2014)
20. M Delcroix, T Yoshioka, A Ogawa, Y Kubo, M Fujimoto, I Nobutaka, K Kinoshita, M Espi, T Hori, T Nakatani, A Nakamura, in *Proceedings of REVERB Challenge Workshop, o2.3*. Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge, (2014)
21. N Epain, T Noohi, C Jin, in *Proceedings of REVERB Challenge Workshop, p2.7*. Sparse recovery method for dereverberation, (2014)
22. X Feng, K Kumatani, J McDonough, in *Proceedings of REVERB Challenge Workshop, p1.9*. The CMU-MIT REVERB challenge 2014 system: description and results, (2014)
23. JT Geiger, E Marchi, BW Schuller, G Rigoll, in *Proceedings of REVERB Challenge Workshop, p1.6*. The TUM system for the REVERB challenge: recognition of reverberated speech using multi-channel correlation shaping dereverberation and blstm recurrent neural networks, (2014)
24. DR Gonzalez, SC Arias, JRC de Lara, in *Proceedings of REVERB Challenge Workshop, p2.2*. Single channel speech enhancement based on zero phase transformation in reverberated environments, (2014)
25. H-G Hirsch, in *Proceedings of REVERB Challenge Workshop, p2.9*. Extraction of robust features by combining noise reduction and FDLP for the recognition of noisy speech signals in hands-free mode, (2014)
26. H Kallasjoki, J Gemmeke, K Palomäki, A Beeston, G Brown, in *Proceedings of REVERB Challenge Workshop, o2.1*. Recognition of reverberant speech by missing data imputation and NMF feature enhancement, (2014)
27. K Kondo, in *Proceedings of REVERB Challenge Workshop, p2.4*. A computationally restrained and single-channel blind dereverberation method utilizing iterative spectral modifications, (2014)
28. YR Leng, J Dennis, WZT Ng, TH Dat, in *Proceedings of REVERB Challenge Workshop, p2.11*. PBF-GSC beamforming for ASR and speech enhancement in reverberant environments, (2014)
29. N López, G Richard, Y Grenier, I Bourmeyster, in *Proceedings of REVERB Challenge Workshop, p2.3*. Reverberation suppression based on sparse linear prediction in noisy environments, (2014)
30. M Mimura, S Sakai, T Kawahara, in *Proceedings of REVERB Challenge Workshop, p1.8*. Reverberant speech recognition combining deep neural networks and deep autoencoders, (2014)
31. V Mitra, W Wang, Y Lei, A Kathol, G Sivaraman, C Espy-Wilson, in *Proceedings of REVERB Challenge Workshop, p2.5*. Robust features and system fusion for reverberation-robust speech recognition, (2014)
32. M Moshirynia, F Razzazi, A Haghbin, in *Proceedings of REVERB Challenge Workshop, p1.2*. A speech dereverberation method using adaptive sparse dictionary learning, (2014)
33. K Ohtani, T Komatsu, T Nishino, K Takeda, in *Proceedings of REVERB Challenge Workshop, p1.5*. Adaptive dereverberation method based on complementary Wiener filter and modulation transfer function, (2014)

34. K Palomäki, in *Proceedings of REVERB Challenge Workshop, p1.10*. Reverberation robust speech recognition by matching distributions of spectrally and temporally decorrelated features, (2014)
35. PP Parada, D Sharma, PA Naylor, T van Waterschoot, in *Proceedings of REVERB Challenge Workshop, p1.4*. Single-channel reverberant speech recognition using C50 estimation, (2014)
36. Y Tachioka, T Narita, FJ Weninger, S Watanabe, in *Proceedings of REVERB Challenge Workshop, p1.3*. Dual system combination approach for various reverberant environments with dereverberation techniques, (2014)
37. J Veras, T Prego, A Lima, T Ferreira, S Netto, in *Proceedings of REVERB Challenge Workshop, p1.7*. Speech quality enhancement based on spectral subtraction, (2014)
38. X Wang, Y Guo, X Yang, Q Fu, Y Yan, in *Proceedings of REVERB Challenge Workshop, p2.1*. Acoustic scene aware dereverberation using 2-channel spectral enhancement for REVERB challenge, (2014)
39. FJ Weninger, S Watanabe, J Le Roux, J Hershey, Y Tachioka, JT Geiger, BW Schuller, G Rigoll, in *Proceedings of REVERB Challenge Workshop, o1.1*. The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement, (2014)
40. S Wisdom, T Powers, L Atlas, J Pitton, in *Proceedings of REVERB Challenge Workshop, p1.11*. Enhancement of reverberant and noisy speech by extending its coherence, (2014)
41. X Xiao, Z Shengkui, DHH Nguyen, Z Xionghu, D Jones, E-S Chng, H Li, in *Proceedings of REVERB Challenge Workshop, o2.2*. The NTU-ADSC systems for reverberation challenge 2014, (2014)
42. F Xiong, N Moritz, R Rehr, J Anemueller, B Meyer, T Gerkmann, S Doclo, S Goetze, in *Proceedings of REVERB Challenge Workshop, p2.8*. Robust ASR in reverberant environments using temporal cepstrum smoothing for speech enhancement and an amplitude modulation filterbank for feature extraction, (2014)
43. M Yu, F Soong, in *Proceedings of REVERB Challenge Workshop, p2.6*. Speech dereverberation by constrained and regularized multi-channel spectral decomposition: evaluated on REVERB challenge, (2014)
44. K Lebart, JM Boucher, PN Denbigh, A new method based on spectral subtraction for speech de-reverberation. Acta Acoustica. **87**, 359–366 (2001)
45. P Swietojanski, A Ghoshal, S Renals, in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition, (2013), pp. 285–290
46. M Seltzer, D Yu, Y Wang, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An investigation of deep neural networks for noise robust speech recognition, (2013), pp. 7398–7402
47. T Zernicki, M Bartkowiak, M Doma, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Enhanced coding of high-frequency tonal components in MPEG-D USAC through joint application of ESBR and sinusoidal modeling, (2011), pp. 501–504
48. LDC Site for The Multi-channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV). https://catalog.ldc.upenn.edu/LDC2014S03. Accessed 13 Jan 2016
49. LDC Site for WSJCAM0. https://catalog.ldc.upenn.edu/LDC95S24. Accessed 13 Jan 2016