REVERB Challenge Speech Enhancement Task

REVERB Challenge Organizers

July 1, 2013

Recently, substantial progress has been made in the field of reverberant speech signal processing, including both single- and multi-channel de-reverberation techniques, and automatic speech recognition (ASR) techniques robust to reverberation. The REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge is organized to evaluate state-of-the-art algorithms and draw new insights regarding potential future research directions. This challenge provides an opportunity to the researchers in the field to carry out a comprehensive evaluation of their methods based on a common database and on common evaluation metrics. This is a multidisciplinary challenge, covering both speech enhancement and speech recognition.

The REVERB challenge consists of a speech enhancement task and an ASR task. This document summarizes the information related to the speech enhancement task. Information common to the ASR task, such as the schedule of the challenge, is available on the challenge website (http://reverb2014.dereverberation.com). The information described here is as of July 1, 2013 and subject to change. Please check the challenge website for the latest information. If you have any questions, please contact the organizers at REVERB-challenge@lab.ntt.co.jp.

1.1 Stages of the Challenge

The challenge consist of two stages: the development stage and the evaluation stage. At the initial stage of the challenge, i.e., the development stage, participants are provided with a training set, a development test set, and an evaluation toolkit. The development test set is a collection of degraded (i.e., noisy and reverberant) speech files that have to be enhanced by participants. The enhancement results are scored with the provided evaluation toolkit. At the development stage, participants are supposed to develop their own speech enhancement systems and optimize their systems to the provided development test set. The training set is a collection of clean speech files and provided mainly for use in the ASR task. But participants in the speech enhancement task are also allowed to use this data set, for example, to develop a reference speech model. A set of script files that generates degraded speech files from the training set is also provided. The resultant degraded training set is called a multi-condition training set and may be used for enhancement.

At the second stage of the challenge, i.e., the evaluation stage, participants will be provided an (final) evaluation test set and asked to evaluate their systems using this test set. Since the period of the evaluation stage is very short, participants are well-advised to optimize their systems at the development stage. This two-stage scheme is taken to enable assessment of "out-of-box performance".

1.2   Data Overview

The data provided consist of a training set, a development test set, and a (final) evaluation test set. The evaluation test set will be made available a couple of weeks before the result submission deadline. Before distribution of the evaluation test set, i.e., at the development stage, participants are supposed to develop their systems based on the development test set (and the training set).

The training set consists of the clean WSJCAM0 training set [1] and a multi-condition training set, which is generated from the clean WSJCAM0 training data by convolving these clean utterances with measured room impulse responses and adding recorded background noise.

The development test set and the evaluation test set each consists of two different parts, namely:
1) SimData – This consists of artificially distorted versions of utterances taken from the WSJCAM0 corpus [1]. These data were created by convolving room impulse responses (RIRs) measured in different rooms with the clean WSJCAM0 signals. Recorded background noise was added to the reverberated data at a fixed signal-to-noise ratio (SNR).
2) RealData – This consists of utterances taken from the MC-WSJ-AV corpus [2]. These data were recorded in a noisy and reverberant room.

The SimData test set captures a broad range of reverberation conditions and thus enables to evaluate the robustness of the approaches in different reverberation conditions. The RealData test set aims at evaluating the robustness of the approaches against variations that are not reproducible by simulation for a certain reverberation condition. Both the development test set and the evaluation test set contain both SimData and RealData. Therefore, participants will be eventually provided with four types of test sets: Development-SimData, Development-RealData, Evaluation-SimData, and Evaluation-RealData.

All reverberant utterances will be provided as 1-channel, 2-channel, and 8-channel recordings. The structure of the evaluation test set will be similar to that of the development test set. Further information, including how to obtain the data sets, can be found on the challenge website (http://reverb2014.dereverberation.com/data.html).

The data sets can be downloaded from the LDC (Linguistic Data Consortium) for free of charge, provided that the data will be used only for this challenge. You will be obtaining the following five materials from LDC, two of which are the most relevant to the speech enhancement task.
- REVERB_WSJCAM0_dt: development set of SimData
- MC_WSJ_AV_Dev: development set of RealData
- wsjcam0: clean training data (mainly used for ASR acoustic model training)
- REVERB_WSJCAM0_tr: multi-condition training data (mainly used for ASR acoustic model training)
- WSJ0_LangMod_REVERB: WSJ language model needed for ASR

In addition, the following two evaluation sets will be available at the beginning of the evaluation stage.

- REVERB_WSJCAM0_et: evaluation set of SimData
- MC_WSJ_AV_Eval: evaluation set of RealData

Please see http://reverb2014.dereverberation.com/data.html for how to download the data.

1.3   Guidelines

To enable fair and reasonable comparison and to ensure that the tasks reproduce reasonably realistic condition, the organizers have set up guidelines for the challenge. They are described on the challenge website (http://reverb2014.dereverberation.com/instructions.html). Participants are expected to follow the guidelines. The guidelines define what information participants can exploit and what information is not allowed to use. These guidelines represent the minimum requirements for making the comparison fair and reasonable. Since one of the main goals of the REVERB challenge is to inspire and evaluate diverse ideas for speech enhancement in reverberant environments, the organizers have tried to minimize the rules for the challenge. These guidelines are applied to both the speech enhancement and ASR tasks. Therefore, some of them are not very relevant to the speech enhancement task. However, participants in the speech enhancement task are expected to respect all the guidelines because some participants may take part in both the speech enhancement and ASR tasks. Please see the challenge website for further details.

1.4   Evaluation for Speech Enhancement

The speech enhancement task aims to let researchers from various fields take part in the challenge. The guidelines shown on the challenge website and the evaluation schemes described below are designed considering two important aspects. Firstly, applications of reverberant speech enhancement are diverse, ranging from hearing aids to automatic speech recognition. Secondly, a universally accepted set of objective quality measures has not been fully established for evaluating reverberant speech enhancement algorithms. Therefore, we have decided to perform both objective and subjective evaluation and use several different objective measures. This means that it is not intended to determine a champion in this task. Rather the goal is to reveal relative merits and demerits of different approaches and also to elucidate the characteristics of each objective quality measure, which will be hopefully facilitating future research and development of reverberant speech enhancement algorithms.

The details of the objective measures used and the subjective evaluation procedure that will be taken are described in the following.

1.4.1   Objective evaluation

The objective measures are split into mandatory ones and optional ones. The mandatory objective measures include ones that have been used for evaluation of reverberant speech enhancement algorithms in the literature. Specifically, the following measures are used.

1. Cepstrum distance: The cepstrum distance (CD) is based on the discrepancy between target and reference signals. For each test utterance, the corresponding clean signal is used as the reference. Thus, the CD is used only for SimData. The CD is calculated as per [3]. To ignore the impact of coloration to the extent possible, cepstral mean

normalization is applied before calculating the distances. Smaller values are assumed to indicate better speech quality.

2. Log likelihood ratio: The log likelihood ratio (LLR) is based on the discrepancy between target and reference signals. For each test utterance, the corresponding clean signal is used as the reference. Thus, the LLR is used only for SimData. The LLR is calculated as per [3]. Smaller values are assumed to indicate better speech quality.

3. Frequency-weighted segmental SNR: The frequency-weighted segmental SNR (FWSegSNR) is based on the discrepancy between target and reference signals. For each test utterance, the corresponding clean signal is used as the reference. Thus, the FWSegSNR is used only for SimData. The FWSegSNR is calculated as per [3]. In the implementation provided, critical band analysis is performed with the mel-frequency filterbank that is often used for ASR. The FWSegSNR is used instead of the segmental SNR, which is often used in the literature, considering the finding in [3] that this measure is more highly correlated with the perceptual speech quality. Larger values are assumed to indicate better speech quality.

4. Speech-to-reverberation modulation energy ratio: The speech-to-reverberation modulation energy ratio (SRMR) can be calculated only from target signals. Thus, the SRNR scores are used for both SimData and RealData. The SRMR scores are calculated as per [4] by using tools provided by the authors of [4]. Larger values are assumed to indicate better speech quality.

5. Wall-clock time: The average wall-clock time spent for processing test utterances is expected to be submitted along with the other quality measures. Since each participant may use different computational environments, direct comparison of runtime is impossible. Therefore, at the beginning of the evaluation stage, a reference enhancement code will be released, and participants are asked to submit the wall-clock time spent for running the reference code.

In addition, we recommend submitting the results obtained with the following optional measures.

6. Word error rate: Participants in the speech enhancement task are strongly encouraged to take part in the ASR task by using their enhancement algorithms as a front-end to the ASR baseline system provided by the challenge. Since the word error rate (WER) can be calculated only from target signals, the WER can be used for both SimData and RealData.

7. PESQ: The enhanced speech signals can also be evaluated in terms of PESQ (Percepsutal Evaluation of Speech Quality as defined by ITU-T Recommendation P.862.2) and the PESQ scores can be submitted along with the numbers of the above described measures. Since PESQ requires reference signals, this can be used only for SimData. Larger values are assumed to indicate better speech quality.

The final result for each measure is calculated by averaging the numbers for individual utterances (i.e., audio files). Since the CD, LLR, and FWSegSNR are calculated for each short time frame, both utterance-level means and utterance-level medians are used for these measures. All numbers should be calculated with the evaluation tools provided by the challenge. The tools are available at

http://reverb2014.dereverberation.com/download.html. Please see the README file included in the distribution for how to use the tools.

1.4.2 Subjective evaluation

As part of this challenge a web-based subjective evaluation will be conducted. Both the challenge participants as well other people are invited to take part in the subjective evaluation. A MUSRHA test will be conducted focusing on perceptual attributes such as perceived distance and overall speech quality. As a reference, the clean signal will be used from the SimData and the close talking recording will be used from the RealData. The listener is asked to listen to the examples using headphones (the type of headphones needs to be specified). The web site for the subjective evaluation will open when the final evaluation test set is released.

1.5 Call for Participation

Reverberation is one of the major factors that severely degrade the audible quality of speech signals and the ASR performance. On one hand, despite some pioneering efforts, enhancement and automatic recognition of reverberant speech has been a very challenging task and yet to be fully solved. On the other hand, research on reverberant speech processing has undoubtedly achieved significant progress in both speech enhancement and recognition fields in recent years. Some of these novel techniques seem ready to be evaluated for real-world speech enhancement and speech recognition applications. Thus, the need for a common evaluation framework seems increasing to enable fair and reasonable comparison of different approaches and provide a clear understanding of the state of the art. This is why we have started organizing this challenge. Therefore, your participation is welcomed and is definitely invaluable to the success of this challenge. The organizers are hoping that this challenge will be providing a better understanding of the state of the art, new insights into the problem, and promising directions of future research.

1.6 References

[1] T. Robinson, J. Fransen, D. Pye and J. Foote and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition", In Proc. ICASSP 95, 81-84, 1995.
[2] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments", In Proc. ASRU, 357-362, 2005.
[3] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE T-ASLP, 16(1), 229-238, 2008.
[4] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," IEEE T-ASLP, 18(7), 1766-1774, 2010.

1.7 Appendix: Database directory trees

1.7.1 MC-WSJ-AV database (development test set, RealData)

```
.
|-- audio
```

```
|    '-- stat
|    |-- T10
|    |    |-- array1
|    |    |-- array2
|    |    |-- headset1
|    |    '-- lapel1
|    |-- T6
|    |    |-- array1
|    |    |-- array2
|    |    |-- headset1
|    |    '-- lapel1
|    |-- T7
|    |    |-- array1
|    |    |-- array2
|    |    |-- headset1
|    |    '-- lapel1
|    |-- T8
|    |    |-- array1
|    |    |-- array2
|    |    |-- headset1
|    |    '-- lapel1
|    '-- T9
|         |-- array1
|         |-- array2
|         |-- headset1
|         '-- lapel1
|-- etc
|    '-- sentencelocation
'-- mlf
'-- WSJ.mlf
30 directories
```

1.7.2   REVERB_WSJCAM0 database (development test set, SimData)

```
.
'—data
    |-- cln_test
    |    '-- primary_microphone
    |         '-- si_dt
    |              |-- c31
    |              |-- ...
    |              '-- c49
    |-- far_test
```

```
|    '-- primary_microphone
|        '-- si_dt
|            |-- c31
|            |-- ...
|            '-- c49
'-- near_test
    '-- primary_microphone
        '-- si_dt
        |-- c31
        |-- ...
        '-- c49
71 directories
```

1.7.3   WSJCAM0 database (clean training set)

```
.
|-- data
|   |-- primary_microphone
|   |   |-- doc
|   |   |-- etc
|   |   |-- si_dt
|   |   |   |-- c31
|   |   |   |-- ...
|   |   |   '-- c49
|   |   |-- si_et_1
|   |   |   |-- c30
|   |   |   |-- ...
|   |   |   '-- c4a
|   |   '-- si_tr
|   |       |-- c02
|   |       |-- ...
|   |       '-- c2l
|   '-- secondary_microphone
|       |-- si_dt
|       |-- si_et_2
|       '-- si_tr
'docs
264 directories
```

1.7.4   REVERB_WSJCAM0_tr database (multi-condition training set)

```
.
'--data
   '-- mc_train
```

```
    '-- primary_microphone
        '-- si_tr
            |-- c02
            |-- ...
            '-- c2l
96 directories
```