Readme for REHASP 0.5: The REpeated HArvard Sentence Prompts corpus version 0.5

Gustav Eje Henter, Thomas Merritt, Matt Shannon, Catherine Mayo, and Simon King

1 Overview

The REHASP corpus contains sentence prompts repeated several times by a single talker. These recordings can be used to investigate the acoustic variability of natural speech. Version 0.5 of the corpus contains 30 different sentence prompts, each repeated 40 times, for a total of 1200 recordings, all freely available under a permissive license. A larger corpus is in preparation.

This document provides meta-information about the REHASP 0.5 database release, including its creation, its contents, and its use. The document is laid out as follows:

2 Availability

The corpus is permanently hosted at Edinburgh DataShare, curated by the University of Edinburgh Information Services. There are two versions: the full corpus, provided in the compressed file `rehasp_0.5.zip` (approximately 2 GiB uncompressed) and a reduced version `rehasp_0.5_16k_std.zip` (approximately 80 MiB uncompressed) which omits the unprocessed 96 kHz recordings.

The audio data is licensed under the Creative Commons Attribution 4.0 International License. Full details of the license are given in `License.txt`. For attribution it is suggested, but not mandated, to cite the scientific publication [1], for instance as:
G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech,* 2014.

3 Contents

The database release contains the following key files and directories:

No.	File or folder name	Description
1	`16k_std/`	Audio folder (16 kHz)
2	`96k/`	Audio folder (96 kHz; full release only)
3	`itu-g191/`	Placeholder for ITU G.191 tools
4	`License.txt`	License terms
5	`Pipeline`	Bash processing commands
6	`prompts/`	Text prompts folder
7	`Readme.pdf`	This document (Adobe pdf format)
8	`Readme.txt`	This document (raw text format)

(A few files of lesser importance have been omitted from this listing.)

4 Prompts

The prompts were adapted from the Harvard sentences [2]. These are sets of ten sentences each, selected to be approximately phonetically balanced within each set. For the REHASP 0.5 corpus, 3 sentence sets suitable for British and US English speakers were selected, namely Harvard sentences 1–10, 31–40, and 71–80.

An example sentence is "Rice is often served in round bowls" (Harvard sentence 5). A full list of the 30 sentences selected is provided in the `prompts` directory of the database.

5 Talker

The talker used for the recordings is "Lucy," a female native speaker of British English in her 20s with a mild British midlands accent (non-RP). Lucy has a slightly breathy voice, meaning that vocoding the recordings typically introduces some quality loss.

Note: The same talker has also recorded other material, including read and spontaneous speech data suitable for training or adapting parametric speech synthesis systems. Please contact Rasmus Dall at r.dall@sms.ed.ac.uk to inquire about access to these other materials.

6 Recording conditions

The recordings were made in the hemi-anechoic chamber at the Informatics Forum at the University of Edinburgh (10 Crichton Street, Edinburgh, EH8 9AB) using two microphones: a DPA 4035 headset microphone and a Sennheiser MKH 800 p48 microphone on a stand with a plosive shield. The recording software operated at 96 kHz, 16 bits. All recordings took place on Friday 2014-02-07 except as indicated in the below table:

Date	2014-02-10									2014-03-07
Harvard sentence	1	5	6	7	9	34	38	73	77	40
Repetition number	2	9	2	2	7	2	3	3	5	20

At the start of recording, the talker was informed that there would be repeated prompts, and was instructed not to intentionally vary the speaking style between repetitions. Prompts were presented one at a time on a display in the recording chamber, and the talker read each one out loud. To avoid list effects, the prompts were presented in random order, subject to the constraint that the same prompt never could appear twice in a row.

An engineer monitored the recording process to avoid clipping and other problems. After successful reading of a prompt, the engineer manually advanced the recording to the next prompt.

7 Processing

The raw recordings can be found in the `96k/` subfolder of the full database release file `rehasp_0.5.zip`. The recordings are provided as WAV files, exactly as captured and partitioned during the recording session.

The database also provides a second, processed and downsampled, version of the recording data. The purpose of the processing was to standardise and prepare the data for the research in [1]. The majority of the processing used SOX version 14.3.2 (http://sourceforge.net/projects/sox/files/sox/14.3.2/) on a GNU Linux platform. The specific steps performed on each file were, in order:

1. Conversion to mono by discarding the stand microphone channel
2. Endpointing to keep (at most) 100 milliseconds of data before first voice activity (SOX `vad -t 6 -s 0.1 -p 0.1`)
3. Endpointing to keep (at most) 300 milliseconds of data after last voice activity (SOX `reverse vad -t 4 -s 0.1 -p 0.3 reverse`)
4. Intermediate-phase high-pass filter to reject low-frequency interference (SOX `sinc -I -t 20 30`)
5. Downsampling to 16 kHz (SOX `rate -vsI 16k`)
6. Amplitude normalisation to −24 dBov following ITU-T recommendation P.56 [3] (block size 256)

A more complete specification is provided in the `Pipeline` file. The processed recordings can be found in the `16k_std/` subfolder included in the full database release, as well as in the reduced version `rehasp_0.5_16k_std.zip`, which excludes the 96 kHz raw recordings.

8 File format

The recording audio is provided as Microsoft WAV files. The directory structure and file names follow the template
`rehasp_0.5/`*S*`/lucy/rep`*R*`/rehasp_`*S*`_lucy_hvd`*N*`_rep`*R*`.wav`

- *S* is a label specifying the sampling rate and processing applied (either `96k` for the raw recordings, or `16k_std` for the standardised data);
- *N* is a four-digit number identifying the Harvard sentence prompt read (range 0001–0010, 0031–0040, 0071–0080);
- *R* is a three-digit number specifying the repetition number (range 001–040).

The 1200 raw 96 kHz files are in stereo format, with channel 1 corresponding to the headset microphone and channel 2 being the microphone on the stand. The 1200 processed recordings are in mono format and are exclusively based on the headset microphone. All files have a bit depth of 16 bits (signed).

9 Contact information

The corpus was recorded and prepared by Gustav Eje Henter, Thomas Merritt, Matt Shannon, Catherine Mayo, and Simon King for the research paper [1]. All are at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh in the U.K., except for Matt Shannon at the Engineering Department (CUED) at the University of Cambridge, U.K.

The authors can be contacted through e-mail at ghenter@inf.ed.ac.uk (Gustav Eje Henter), t.merritt@ed.ac.uk (Thomas Merritt), sms46@eng.cam.ac.uk (Matt Shannon), catherin@inf.ed.ac.uk (Catherine Mayo), and Simon.King@ed.ac.uk (Simon King). Alternatively, one can write to:

The Centre for Speech Technology Research
Informatics Forum, the University of Edinburgh
10 Crichton Street
Edinburgh
City of Edinburgh
EH8 9AB
UNITED KINGDOM

References

[1] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, 2014.

[2] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE T. Acoust. Speech*, vol. 17, no. 3, pp. 225–246, 1969.

[3] *Objective measurement of active speech level*, ITU Recommendation ITU-T P.56, International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland, March 2011.