NTIA Technical Memorandum 17-523

A Crowdsourced Speech Intelligibility Test that Agrees with, Has Higher Repeatability than, Lab Tests

Stephen D. Voran
Andrew A. Catellier

technical memorandum

U.S. DEPARTMENT OF COMMERCE • National Telecommunications and Information Administration

NTIA Technical Memorandum 17-523

A Crowdsourced Speech Intelligibility Test that Agrees with, Has Higher Repeatability than, Lab Tests

Stephen D. Voran
Andrew A. Catellier

U.S. DEPARTMENT OF COMMERCE

February 2017

DISCLAIMER

Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

PREFACE

The Institute for Telecommunication Sciences has conducted many speech intelligibility tests in support of multiple agencies over the past nine years. These tests have been conducted in highly-controlled laboratory environments, consistent with historical precedent. We recently experimented with crowdsourced speech intelligibility testing using anonymous, self-selecting subjects in uncontrolled environments listening via Internet. The two approaches are nearly opposites in many regards but the results attained are nearly identical. This report explains our motivations, test designs, analyses, and conclusions regarding this fundamentally different way to assess speech intelligibility.

CONTENTS

FIGURES

TABLES

ABBREVIATIONS/ACRONYMS

CMRT	Crowdsourced Modified Rhyme Test
dBA	Decibels, A-Weighted
GMT	Greenwich Mean Time
HIT	Human Intelligence Task
ITU-T	International Telecommunication Union, Telecommunication Standardization Sector
LMRT	Laboratory Modified Rhyme Test
MRT	Modified Rhyme Test
MTurk	Mechanical Turk
QoE	Quality of Experience

EXECUTIVE SUMMARY

Subjective speech intelligibility testing is normally preformed in highly-controlled laboratory conditions. The motivation is to minimize uncontrolled sources of variation so that observed variation can be attributed to the controlled parameters under study in the test. We have followed this paradigm in conducting the Modified Rhyme Test (MRT) for years. Most recently, we have explored a diametrically opposed alternative: the Crowdsourced Modified Rhyme Test (CMRT). For the CMRT we relinquish most control and turn to anonymous, self-selecting listeners using unspecified equipment in uncontrolled environments. These listeners participate via Internet. In exchange for giving up much of the experimental control, we get results from large numbers of listeners quite rapidly and at minimal cost.

The crowdsourcing of subjective Quality of Experience (QoE) tests has been studied for some time. But speech intelligibility testing offers a unique opportunity not found in QoE testing. In speech intelligibility testing, each trial has a correct answer and this allows us to motivate listeners to make their best effort and to evaluate which listeners have been most successful.

The report uses our existing Laboratory Modified Rhyme Test (LMRT) protocol and results as the reference against which CMRT is explored. The specific LMRT used here was designed for 32 listeners and produced 384 MRT trials per condition-under-test. Each listener heard every condition twelve times and every condition was evaluated using the same 384 MRT trials.

Next, the report describes our CMRT platform and protocol. A key component is that we motivate CMRT listeners through a bonus structure, and then evaluate them according to their level of success in the CMRT tasks. We report that the CMRT was completed in less than six hours and, during that time period, 345 different participants completed a total of 86,016 individual CMRT trials or 1536 trials per condition-under-test. This is four times the data of the LMRT, and over 10 times the number of listeners.

Based on careful analysis of LMRT and CMRT data, we arrive at an efficient CMRT design and listener selection rule, driven by the goal of obtaining CMRT results that align with LMRT reference results. That rule is to have two listeners perform each block of CMRT trials, but to retain only the results from the more successful of those two listeners. This process produces CMRT success rates just slightly higher than LMRT success rates. More importantly, for the 56 conditions-under-test considered in this report, CMRT and LMRT produce statistically equivalent results for 55 of them, and the final condition shows only a borderline significant difference.

Through further analysis, we find that CMRT is more repeatable than LMRT. This result stems from the fundamental difference in the LMRT and CMRT test design philosophies. Lab listeners each perform hours of testing to make the visit worthwhile and a relatively small number of listeners will consume the test budget. CMRT listeners are allowed to make much shorter time commitments and we compensate by using large numbers of listeners. This use of large numbers of listeners means that variation due to listeners averages out quickly. This listener variation factor outweighs the reduction in experimental control and causes CMRT to be more repeatable than LMRT.

Compared to a typical laboratory test, the CMRT can save thousands of U.S. dollars, weeks of testing, and eliminate the need for a laboratory environment. But our laboratories remain vital assets in our research program. They provide the tightly-controlled environments that are required for detecting small changes near the threshold of audibility, and are also sometimes needed for QoE testing.

A CROWDSOURCED SPEECH INTELLIGIBILITY TEST THAT AGREES WITH, HAS HIGHER REPEATABILITY THAN, LAB TESTS

Stephen D. Voran and Andrew A. Catellier[1]

Crowdsourcing of subjective speech, audio, and video quality of experience (QoE) tests has received much interest and study, but crowdsourcing of speech intelligibility testing has not. We hypothesize that speech intelligibility tests offer a unique crowdsourcing opportunity because, unlike QoE testing, each trial has a correct answer. That allows us to motivate and evaluate listeners. We describe the design, implementation, and analysis of a Crowdsourced Modified Rhyme Test (CMRT) that replicates our recent Laboratory MRT (LMRT) work. Our results show that CMRT results are more repeatable than LMRT results, CMRT repeats LMRT better than LMRT repeats itself, and application of a simple listener selection rule produces per-condition CMRT results that almost exactly agree with reference LMRT results.

Keywords: Crowdsource, modified rhyme test, MRT, speech intelligibility, subjective test

1. INTRODUCTION

The Institute for Telecommunication Sciences (ITS) has conducted numerous subjective speech tests addressing quality, intelligibility, emotion detection, talker identity, and more. Examples of our published results can be found in [1]–[11]. The motivations for these tests have been diverse, but the test paradigm has been largely consistent—a small number of carefully selected listeners (typically 24 to 48) perform very specific listening and scoring tasks in highly controlled laboratory environments. Crowdsourcing inverts this paradigm—a large number of self-selecting listeners (hundreds or thousands) perform the tasks remotely (via Internet) in completely uncontrolled environments.

Crowdsourcing offers large sample sizes without the financial and logistical constraints associated with procuring and equipping a laboratory and recruiting and processing individual subjects through test protocols at a specific laboratory location. Crowdsourcing also means giving up nearly every element of control over listeners, listening equipment, and listening environment, and this loss of control runs counter to the basic premise of laboratory work. On the other hand, for some applications, the aggregation of uncontrolled listeners, equipment, and conditions inherent in crowdsourcing could make its results more relevant than laboratory results.

Internet-based speech quality testing saw early use in [12] and on-line workshop discussions (captured in the printed workshop proceedings) highlighted opportunities associated with

[1] The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

crowdsourcing and as well as strong resistance to the loss of laboratory control. Optimism won out and researchers began exploring the use of an anonymous Internet-connected crowd to perform quality of experience (QoE) testing for speech, audio, and video. Example advances and specific applications can be found in [13]–[17] while [18] and [19] contain broader studies with extensive bibliographies.

In recent years, our audio quality research efforts at ITS have been largely focused on evaluating the speech intelligibly of deployed and emerging communication systems, often for the public safety user community [8]–[11]. Quantifying the different levels of *speech quality* can be useful in commercial telecommunications as these levels may drive consumer choices and satisfaction. But in public safety communications, *speech intelligibility* may literally be a matter of life and death. Public safety users often must efficiently communicate critical information and successful communications can improve safety, minimize property loss, or even save lives. In addition, public safety users commonly find themselves needing to communicate in the presence of high noise levels (e.g., sirens, alarms, firefighting equipment, and rescue equipment). Thus, for public safety users, speech intelligibility, not speech quality, becomes a critical parameter as it quantifies how well a system preserves speech information, even in the context of noise. Following the precedent set by the National Fire Protection Association, we adopted the Modified Rhyme Test (MRT) [20] protocol for subjectively testing the speech intelligibility of communication systems. That protocol is summarized later in this report.

We hypothesized that the MRT could be successfully crowdsourced because each trial has a correct answer. This provides the opportunity to evaluate listener efficacy and to motivate, reward, and select listeners. Note that this opportunity is not present in QoE testing as there is no single “correct” opinion of quality. We are not aware of any previous crowdsourced implementation of the modified rhyme test (MRT) speech intelligibility test protocol.

Given our hypothesis and our access to Laboratory MRT (LMRT) data to use as references, we elected to design, implement, and analyze a crowdsourced MRT (CMRT) to replicate our most recent LMRT. This paper details CMRT design, implementation and results. First, we briefly summarize the LMRT that serves as a reference for our work. Next, we detail the design and implementation of the CMRT. We then report CMRT timing, cost, and listener participation. Finally, and perhaps most importantly, we analyze the speech intelligibility data collected by CMRT and compare it with the analogous data collected in the reference LMRT. We introduce listener selection rules and we find that CMRT and LMRT data are extremely closely aligned, and that CMRT results have higher repeatability than LMRT results.

2. LABORATORY MODIFIED RHYME TEST

The MRT protocol specifies 50 lists and each contains six English language words with the phonetic pattern consonant-vowel-consonant. In each list, the six words differ only in the leading or trailing consonant. An MRT trial consists of the presentation of one word in a carrier phrase (e.g., “Please select the word bit.”). The listener then selects what was heard from six options (e.g., “kit,” “bit,” “fit,” “hit,” “wit,” and “sit”) on a graphical interface. If the correct word is selected a trial is considered to be a success. Success rates are then translated to speech intelligibility scores by subtracting the success rate for guessing ($\frac{1}{6}$) and then rescaling to the interval [0,1].

We recently completed a large MRT in our laboratories [11] and it provides the reference data for our CMRT investigation. This balanced LMRT was designed for 32 listeners, each performing 6 sessions and each session containing 336 trials. Each session was dedicated to a single noise type and covered 28 codec modes using 12 trials per mode. The result was 384 MRT trials per codec mode. MRT recordings came from two female and two male talkers (96 each.) Every codec mode was tested with the same 384 MRT trials but each listener heard different pairings of codec mode and MRT source.

We recruited 32 listeners (and 4 alternates) employed in public-safety related occupations (e.g., dispatch, fire service, law enforcement, paramedic). Females made up 25% of the listeners and the median estimated age decade was the forties. The LMRT was conducted in sound-isolated chambers where noise measured below 26.5 dBA. One listener occupied each chamber. The listening instrument was a professional grade personal monitor speaker. Each listener was encouraged to adjust the volume to preferred listening level. Full details and results are provided in [11].

3. CROWDSOURCED MODIFIED RHYME TEST

We designed this initial CMRT to parallel our recent LMRT to the extent possible. We elected to replicate two of the six LMRT noise environments in the CMRT: “quiet” and “saw” (highest and lowest average LMRT scores, respectively). We call a combination of a codec mode and a noise environment a “condition” so there are 56 (28 × 2) conditions in this CMRT.

We chose Amazon Mechanical Turk (MTurk) as our CMRT platform. This on-line outsourcing service gave us access to over 500,000 potential workers (listeners in our case) worldwide and provided standard interfaces for interacting with them and for paying base wages and bonuses. The basic MTurk work unit is a human intelligence task (HIT). We defined a HIT to contain 56 MRT trials—one trial from each condition, taken directly from the LMRT playlists. Replication of LMRT for the two noise environments required 384 HITs.

We set payment at $0.56 per HIT (one cent per trial). The average LMRT trial time was 4.1 seconds which translates to an average base pay rate of $8.78 per hour. This safely exceeds the U.S. federal minimum wage of $7.25.

To allow flexibility in this initial investigation of CMRT, we configured the batch of 384 HITS so that four different listeners would complete each HIT, expecting to reduce this number in future CMRT work. Thus we requested 1536 HIT assignments (86,016 trials) from the CMRT listeners.

To motivate and enable listener performance, our instructions included this text: “*We offer incentives to workers who mark the highest number of correct words. We pay a 50 percent bonus to the top 25 percent of workers. Factors that may improve results include: familiarity with English, limited distractions, quiet work environment, and good speakers, earbuds, or headphones.*”

We implemented this bonus payment system on a per-HIT basis. For each HIT, our software counted the number of correct words (0 to 56) marked by each of the four workers and awarded the bonus ($0.28) to the worker with the highest number of correct words.

The interface presented to each worker used JavaScript to enforce the MRT protocol. Listeners clicked a link to begin playback for each trial. As soon as the link was clicked, audio playback began and the six word choices appeared. However, the listener was not allowed to select an option until audio playback completed. No decipherable reference to original filenames or paths was visible in the interface or underlying code. An SHA-256 hash based on filename, HIT number and trial was generated for each trial and stored in a database along with the original filename. The URL for the audio source in each trial was thus constructed using an HTTP GET query using the unique SHA-256 hash for that trial. The server fulfilled the query by consulting the database, fetching the appropriate audio file, and sending the file to the browser.

4. RESULTS

We launched the CMRT at 1:22 PM local time (19:22 GMT) on August 9, 2016. The first listener accepted a HIT 27 seconds later and at one minute after launch 11 HITs had been accepted by listeners.[2] Additional HITs were accepted at a roughly uniform rate (about 78 HITs per 15 minutes) until tapering off around 4.75 hours when only 46 HITs remained available for assignment. The final HIT was accepted 5.3 hours after launch and completed minutes later.

A total of 345 listeners participated: 199 completed a single HIT, 142 completed between 2 and 30 HITS, and the remaining 4 completed 31, 33, 35, and 50 HITs. On average, each listener completed 4.5 HITs. The total cost for wages, bonuses and MTurk fees was US $1160. We collected 86,016 trials in 5.3 hours and 99.84% of the trials produced valid data. Six HITs were superficially corrupted by one or two single empty response fields and three HITs were seriously corrupted by multiple empty response fields.

The operational costs for this CMRT were thousands of dollars lower than those of a typical comparable LMRT. The time required to collect the trials was reduced from weeks to hours. Compared to the specific LMRT described in Section 2 (a specific listener population was recruited and travel expenses were paid), the CMRT saved tens of thousands of dollars and months of time. The software development efforts for CMRT and LMRT are similar. In fact, we were able to repurpose some LMRT design and playlist software for CMRT use. Finally, note that CMRT eliminates the need for a laboratory and that translates to another very substantial cost savings.

4.1 Efficacy of Listeners and Listener Selection Rules

Listener effectiveness in speech intelligibility tests can hinge on many factors including auditory acuity, level of effort, playback equipment, listening environment, and language experience. For the CMRT we cannot control any of these, but we do attempt to influence these in a positive way via the instructions and bonus payments highlighted in Section 3.

The top histogram in Figure 1 shows the success rate for 1536 submitted HITs (3 seriously corrupted HITs are beyond the left limit.) The 32 lab listeners have success rates between 0.778 and 0.869. This range is shown with dashed black lines and is clearly consistent with the histogram mode but the histogram is skewed by a tail of HITs with lower success rates. These lower success rates may be due to any of the factors listed above. The mean and median values are 0.788 and 0.804 respectively and these are slightly lower than the corresponding statistics for the LMRT distribution (0.825 and 0.832 respectively).

But the histogram shows four listeners per HIT, and our goal of reproducing LMRT only requires one listener per HIT. So by design, we now have the luxury of selecting one listener per HIT. This requires that we develop a listener selection rule. The goal of the listener section rule is to select listeners who are willing and able to produce an effort similar to the listeners used in

[2] Our reporting throughout is limited to HITs accepted and successfully completed. We did not track abandoned HITs as abandoned HITs were automatically and immediately made available to other listeners.

LMRT. This means minimizing the tail of low success HITs, leaving a CMRT success rate histogram that more closely approximates the LMRT success histogram.

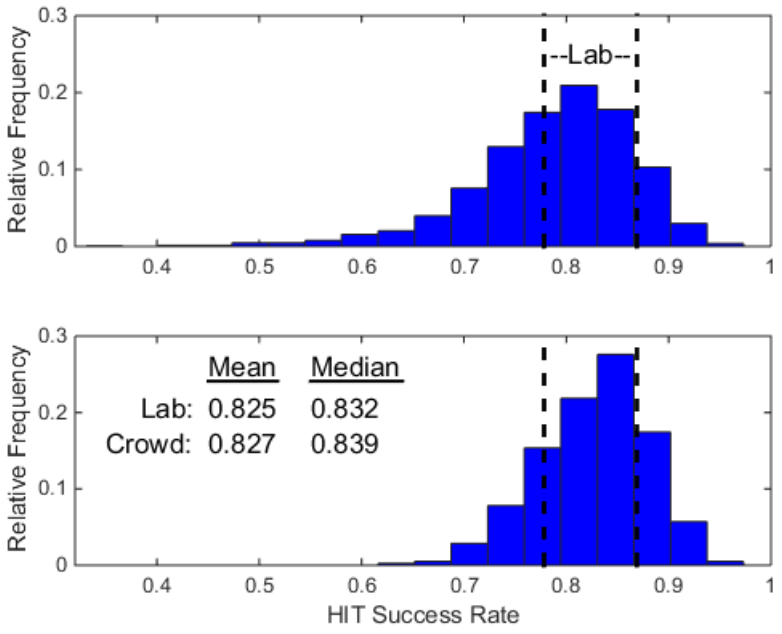

Figure 1. Histograms of CMRT success rates for all 1536 HITs (upper) and for 384 HITs selected by taking the more successful listener of the first two listeners (lower).

We can evaluate listener selection rules by comparing the resulting CMRT data with the LMRT reference data. But a listener selection rule cannot use any LMRT data because in general, none would be available.

We have identified multiple listener selection rules that trim the left tail of the histogram and thus bring the CMRT success statistics into very close alignment with LMRT statistics. One option is to use the first three listeners who complete a HIT, and then select the listener with the median success rate. But a more efficient option is to use only the first two listeners who complete a HIT and select the listener with the higher success rate. This listener selection rule produces the lower histogram in Figure 1. Note that the mean and median success rates are now just slightly above those of LMRT. After selecting results by this rule, 152 unique listeners are represented, and the average number of retained hits per listener is 2.5. This rule requires only two listeners per HIT. This design would reduce the cost of wages, bonuses, and MTurk fees to $580.

We will adopt this listener selection rule (two workers per HIT, use results from the more successful worker) for CMRT designs going forward. As any additional CMRT to LMRT comparisons become available this rule can be further verified or refined as necessary. Because every listener hears each condition exactly once, listener selection is not expected to skew the results for any condition relative to any other condition. As shown below, the retained listeners show a wide range of per-condition success rates that are in good agreement with those found in LMRT. The listener selection rule produces a set of anonymous, remote, listeners operating in

uncontrolled environments for modest pay who are every bit as effective as the listeners we individually recruit and supervise in laboratory tests.

4.2 Crowd-Lab Agreement

We have established that CMRT listeners can be just as successful as LMRT listeners. But the goal of LMRT and CMRT is to evaluate speech intelligibility of conditions under test, not success of listeners. Figure 2 shows the LMRT and CMRT success rates (speech intelligibility) for the 56 conditions in this study. The CMRT results use the more successful of the first two listeners. Thus each data point represents 384 LMRT trials and 384 CMRT trials.

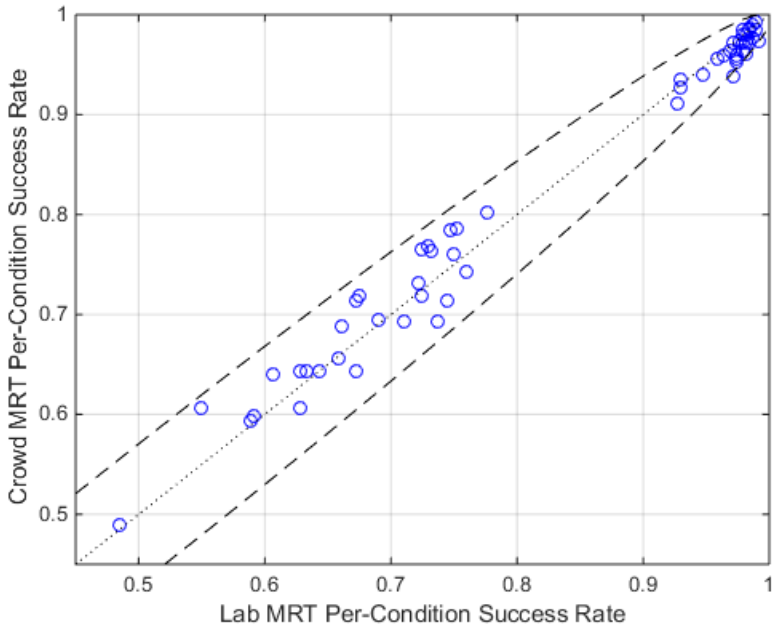

Figure 2. CMRT and LMRT success rates for 56 conditions.

The dashed lines in Figure 2 show the region of statistical equivalence (LMRT serving as reference) at the 95% significance level based on the chi-squared test for independence of categorical data [11], [21]. Fifty-five of the 56 conditions fall inside the regions of statistical equivalence. The final condition is truly a borderline case and had the 360 CMRT successes instead been 362 successes, it would also fall inside the region of equivalence. We conclude that LMRT and CMRT produce very similar per-condition results here.

4.3 Repeatability of Results

Given this outcome it may be tempting to ask if the LMRT results or the CMRT results are "more correct." Of course there is no way to answer this, but we can investigate another critical desirable property of any measurement: repeatability. Recall that LMRT was designed for 32 listeners but 4 alternates were recruited. These alternates became listeners 33-36 and they

repeated the exact trials of listeners 1-4 respectively. These replications allow us calculate 224 samples of the per-condition lab-to-lab repeatability measure,

$$\delta_{LL}(c,u) = \frac{1}{N}\left|n_{c,(u+32)} - n_{c,u}\right|, 1 \le c \le 56, 1 \le u \le 4, \quad (1)$$

where $n_{c,u}$ is the number of successful trials out of the $N = 12$ total trials for condition c and lab listener u. We selected this as the basic data unit for repeatability tests because it is the largest (and hence least noisy) meaningful unit (it addresses a single condition) that is exactly repeated between two listeners. The statistic $\delta_{LL}(c,u)$ compares the second lab listener with the first lab listener on identical trials.

We can calculate 224 perfectly analogous samples of the per-condition crowd-to-crowd repeatability measure:

$$\delta_{CC}(c,u) = \frac{1}{N}\left|m_{c,u,2} - m_{c,u,1}\right|, 1 \le c \le 56, 1 \le u \le 4. \quad (2)$$

Here $m_{c,u,v}$ is the number of successful trials out of a group of $N = 12$ trials, for crowd listener v and condition c. The index u specifies the twelve trials via,

$$m_{c,u,v} = \sum_{h=12(u-1)+1}^{12u} s_{c,h,v}\ , \quad (3)$$

where $s_{c,h,v}$ is the binary success indicator for the trial in HIT h that addresses condition c performed by crowd listener v. The function of (3) is to aggregate from the CMRT results the 12 HITs that contain the 12 trials that were heard in LMRT by listener u. This makes $m_{c,u,v}$ analogous to $n_{c,u}$ for each of the four crowd listeners $v = 1$ to 4. The statistic $\delta_{CC}(c,u)$ compares the second group of crowd listeners on the selected trials with the first group of crowd listeners on the identical trials. This is a *worst-case analysis* of crowd-to-crowd repeatability because *no listener selection rule* has been applied.

We can also calculate 224 samples of the analogous per-condition crowd-to-lab repeatability measure:

$$\delta_{CL}(c,u) = \frac{1}{N}\left|m_{c,u,1} - n_{c,u,}\right|, 1 \le c \le 56, 1 \le u \le 4. \quad (4)$$

The statistic $\delta_{CL}(c,u)$ compares the first crowd listeners with the first lab listener on identical trials.

The distributions of $\delta_{LL}(c,u)$ and $\delta_{CC}(c,u)$ describe the repeatability of the outcomes of groups of 12 trials in LMRT and CMRT while $\delta_{CL}(c,u)$ describes the ability of CMRT to replicate LMRT. These distributions are summarized by statistics in Table 1 and cumulative histograms in Figure 3. These presentations make it clear that CMRT is more repeatable than LMRT.

Table 1. Summary statistics for three MRT repeatability measures. Smaller values indicate higher repeatability.

	Lab-Lab (δ_{LL})	**Crowd-Lab (δ_{CL})**	**Crowd-Crowd (δ_{CC})**
Mean	0.111	0.100	0.085
Standard Deviation	0.128	0.098	0.087
Maximum	0.667	0.500	0.500

We attribute this to the fact that $\delta_{LL}(c,u)$ captures the difference between 2 listeners, but $\delta_{CC}(c,u)$ captures the difference between (up to) 12 listeners and (up to) 12 other listeners, thus some listener-to-listener variation is averaged out before $\delta_{CC}(c,u)$ is computed. This mathematical difference reflects the difference in the LMRT and CMRT test design philosophies. Lab listeners each perform hours of testing to make the visit worthwhile and a relatively small number of listeners will consume the test budget. CMRT listeners are allowed to make much shorter time commitments and we compensate by using large numbers of listeners. Finally, note that comparing $\delta_{CL}(c,u)$ with $\delta_{LL}(c,u)$ shows that CMRT actually reproduces LMRT better than LMRT reproduces itself.

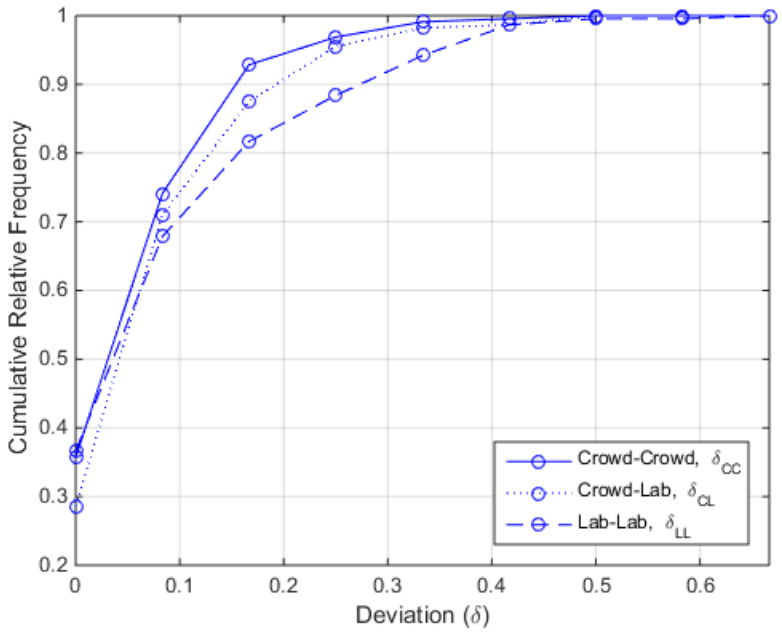

Figure 3. Cumulative histograms for repeatability measures.

5. CONCLUSIONS

CMRT can save significant time and expense and greatly increase listener diversity. CMRT results are more repeatable than LMRT results and CMRT repeats LMRT better than LMRT repeats itself. A simple listener selection rule produces CMRT listener success rates slightly above those of LMRT, and per-condition CMRT results that agree with reference LMRT results for 55 of 56 conditions. We note that it should be straight-forward to extend the CMRT framework to support other forced-choice speech intelligibility test protocols, including those utilizing the Grid Corpus [22] and the method recently recommended by ITU-T [23].

6. REFERENCES

[1] S. Voran, "Listener ratings of speech passbands," in *Proc. 1997 IEEE Workshop on Speech Coding for Telecommunications*, Sept. 1997, pp. 81–82.

[2] S. Voran, "Perception of temporal discontinuity impairments in coded speech - a proposal for objective estimators and some subjective test results," in *Proc. 2nd International Measurement of Speech and Audio Quality in Networks Conference*, Prague, Czech Republic, May 2003, pp. 37–46.

[3] S. Voran, "A basic experiment on time-varying speech quality," in *Proc. 4th International Measurement of Speech and Audio Quality in Networks Conference*, Prague, Czech Republic, June 2005, pp. 51–64.

[4] S. Voran, "Listener detection of talker stress in low-rate coded speech," in Proc. *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, March 2008, pp. 4813–4816.

[5] S. Voran and A. Catellier, "Gradient ascent paired comparison subjective quality testing," in *Proc. First International Workshop on Quality of Multimedia Experience*, QoMEX 2009, July 2009, pp. 133–138.

[6] S. Voran, "Subjective ratings of instantaneous and gradual transitions from narrowband to wideband active speech," in *Proc. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4674 –4677.

[7] S. Voran and A. Catellier, "When should a speech coding quality increase be allowed within a talk-spurt?," in *Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8149–8153.

[8] D. Atkinson and A. Catellier, "Intelligibility of selected radio systems in the presence of fireground noise," Tech. Rep. TR-08-453, NTIA, 2008.

[9] D. Atkinson, S. Voran, and A. Catellier, "Intelligibility of the adaptive multi-rate speech coder in emergency response environments," Tech. Rep. TR-13-493, NTIA, 2012.

[10] D. Atkinson and A. Catellier, "Intelligibility of analog FM and updated P25 radio systems in the presence of fireground noise: Test plan and results," Tech. Rep. TR-13-495, NTIA, 2013.

[11] S. Voran and A. Catellier, "Speech codec intelligibility testing in support of mission-critical voice applications for LTE," Tech. Rep. TR-15-520, NTIA, 2015.

[12] L. Sun and E. Ifeachor, "Subjective and objective speech quality evaluation under bursty losses," in *Proc. International Measurement of Speech and Audio Quality in Networks Conference*, Prague, Czech Republic, Jan. 2002, pp. 25–27.

[13] K. T. Chen, C. C. Wu, Y. C. Chang, and C. L. Lei, “A crowdsourceable QoE evaluation framework for multimedia content,” in *Proc. 7th ACM International Conference on Multimedia*, New York, 2009, pp. 491–500.

[14] M. Wolters, K. Isaac, and S. Renals, “Evaluating speech synthesis intelligibility using Amazon Mechanical Turk,” in *Proc. 7th Speech Synthesis Workshop*, 2010, pp. 136–141.

[15] S. Buchholz and J. Latorre, “Crowdsourcing preference tests, and how to detect cheating,” in *Proc. Interspeech 2011*, 2011, pp. 3053–3056.

[16] F. Ribeiro, D. Florłncio, C. Zhang, and M. Seltzer, “CrowdMOS: An approach for crowdsourcing mean opinion score studies,” in *Proc. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp. 2416–2419.

[17] M. Mandel, “Learning an intelligibility map of individual utterances,” in *Proc. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013.

[18] C. C. Wu, K. T. Chen, Y. C. Chang, and C. L. Lei, “Crowdsourcing multimedia QoE evaluation: A trusted framework,” *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1121–1137, Aug. 2013.

[19] T. Hossfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, “Best practices and recommendations for crowdsourced QoE,” Tech. Rep. 1.0, COST Action IC 1003 Qualinet, 2014.

[20] ANSI/ASA S3.2-2009, “Method for Measuring the Intelligibility of Speech over Communication Systems,” 2009.

[21] R. Hogg and E. Tanis, *Probability and Statistical Inference*, Macmillan, 1977.

[22] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, 2006.

[23] ITU-T Rec. P.807, “Subjective test methodology for assessing speech intelligibility,” 2016.

NTIA FORM 29
(4-80)

U.S. DEPARTMENT OF COMMERCE
NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO.	2. Government Accession No.	3. Recipient's Accession No.
TM-17-523		

4. TITLE AND SUBTITLE	5. Publication Date
A Crowdsourced Speech Intelligibility Test that Agrees with, has Higher Repeatability than, Lab Tests	February 2017
	6. Performing Organization Code NTIA/ITS.P

7. AUTHOR(S)	9. Project/Task/Work Unit No.
Stephen D. Voran and Andrew A. Catellier	3103011-300
8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305	10. Contract/Grant Number.

11. Sponsoring Organization Name and Address	12. Type of Report and Period Covered
National Telecommunications & Information Administration Herbert C. Hoover Building 14th & Constitution Ave., NW Washington, DC 20230	Technical Memorandum

14. SUPPLEMENTARY NOTES

15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

Crowdsourcing of subjective speech, audio, and video quality of experience (QoE) tests has received much interest and study, but crowdsourcing of speech intelligibility testing has not. We hypothesize that speech intelligibility tests offer a unique crowdsourcing opportunity because, unlike QoE testing, each trial has a correct answer. That allows us to motivate and evaluate listeners. We describe the design, implementation, and analysis of a Crowdsourced Modified Rhyme Test (CMRT) that replicates our recent Laboratory MRT (LMRT) work. Our results show that CMRT results are more repeatable than LMRT results, CMRT repeats LMRT better than LMRT repeats itself, and application of a simple listener selection rule produces per-condition CMRT results that almost exactly agree with reference LMRT results.

16. Key Words (Alphabetical order, separated by semicolons)

Crowdsource, modified rhyme test, MRT, speech intelligibility, subjective test

17. AVAILABILITY STATEMENT	18. Security Class. (This report)	20. Number of pages
☒ UNLIMITED.	Unclassified	29
☐ FOR OFFICIAL DISTRIBUTION.	19. Security Class. (This page) Unclassified	21. Price: n/a

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail itsinfo@ntia.doc.gov.