Journal of Communication Disorders 53 (2015) 70–83

Contents lists available at ScienceDirect

Journal of Communication Disorders

Online crowdsourcing for efficient rating of speech: A validation study

Tara McAllister Byun [*], Peter F. Halpin, Daniel Szeredi

New York University, New York, NY, USA

ARTICLE INFO

Article history:
Received 4 July 2014
Received in revised form 30 October 2014
Accepted 23 November 2014
Available online 15 December 2014

Keywords:
Crowdsourcing
Speech rating
Research methods
Speech sound disorders
Speech perception

ABSTRACT

Blinded listener ratings are essential for valid assessment of interventions for speech disorders, but collecting these ratings can be time-intensive and costly. This study evaluated the validity of speech ratings obtained through online crowdsourcing, a potentially more efficient approach. 100 words from children with /r/ misarticulation were electronically presented for binary rating by 35 phonetically trained listeners and 205 naïve listeners recruited through the Amazon Mechanical Turk (AMT) crowdsourcing platform. Bootstrapping was used to compare different-sized samples of AMT listeners against a “gold standard” (mode across all trained listeners) and an “industry standard” (mode across bootstrapped samples of three trained listeners). There was strong overall agreement between trained and AMT listeners. The “industry standard” level of performance was matched by bootstrapped samples with $n = 9$ AMT listeners. These results support the hypothesis that valid ratings of speech data can be obtained in an efficient manner through AMT. Researchers in communication disorders could benefit from increased awareness of this method.

Learning outcomes: Readers will be able to (a) discuss advantages and disadvantages of data collection through the crowdsourcing platform Amazon Mechanical Turk (AMT), (b) describe the results of a validity study comparing samples of AMT listeners versus phonetically trained listeners in a speech-rating task.

1. Introduction

To study the efficacy of interventions for speech sound disorder, researchers must identify a valid, reliable method for measuring changes in speech production accuracy or intelligibility over time. Acoustic and other instrumental measures are an indispensable part of this process, but from a clinical standpoint, it is most important to know whether treatment yields a meaningful change in human listeners’ perception of speech. Given the potential for bias when raters are familiar with the participants or purpose of a study, it is essential to collect these ratings in a blinded, randomized fashion. However, obtaining these ratings can pose a major challenge for speech researchers.

The conventional approach to obtaining ratings of speech data is a multi-step process. First, it is necessary to identify potential raters, typically certified clinicians (e.g., McAllister Byun & Hitchcock, 2012) or students in speech-language pathology (e.g., Maas & Farinella, 2012). Potential raters are generally required to complete some number of practice or training trials, followed by a test to determine their eligibility to participate. If an individual meets a predetermined

[*] Corresponding author at: Department of Communicative Sciences and Disorders, New York University, 665 Broadway, Room 914, New York, NY 10012, USA. Tel.: +1 212 992 9445; fax: +1 212 995 4356.
E-mail address: tara.byun@nyu.edu (T. McAllister Byun).

http://dx.doi.org/10.1016/j.jcomdis.2014.11.003
0021-9924/© 2014 Elsevier Inc. All rights reserved.

threshold of accuracy or agreement with a ‘‘gold standard’’ rater, he/she is invited to complete the primary rating task. After ratings are collected, some or all of the responses provided by one listener must be compared against another listener’s responses in order to calculate interrater reliability. If agreement between raters falls below a critical threshold, it may be necessary to exclude raters, or to elicit ratings again after providing additional training.

This conventional approach to speech sound rating is well-established in the literature, having produced many examples in which interrater agreement exceeds 80% (Shriberg et al., 2010; Shriberg & Lof, 1991). However, its primary drawback is that it is time-consuming and requires intensive effort on the part of the experimenter as well as the rater. Because most commercially available programs for stimulus presentation and response recording require manual installation of proprietary software, the process typically involves multiple points of in-person contact between the raters and the research team. Researchers may struggle to find an adequate number of listeners, and excluding listeners due to low interrater agreement can represent a major setback. Moreover, obtaining ratings can incur significant costs for the researcher. While it is desirable to use the expert judgment of certified clinicians, many researchers are unable to offer compensation in line with the typical pay rate of speech-language pathologists, who earn a median hourly wage of $50.00 (American Speech-Language-Hearing Association, 2012). Facing these challenges, researchers may resort to non-optimal methods, such as using the authors or other study personnel as data raters. When listeners are familiar with the experimental design and participants, there is a significant risk that their ratings will reflect some influence of preexisting bias.

A novel solution to the longstanding problem of obtaining speech ratings has arisen in connection with recent technological innovations in the area of online crowdsourcing. In crowdsourcing, a task or problem is assigned not to a small number of specialists, but to a large number of non-experts recruited through online channels. Taken individually, these non-specialists do not display expert levels of performance, but in the aggregate, their responses generally converge with those assigned by experts (Ipeirotis, Provost, Sheng, & Wang, 2014). Crowdsourcing methods have succeeded in solving remarkably complex problems, as in a case where untrained users playing an online game arrived at an accurate model of a protein structure that had previously eluded trained scientists (Khatib et al., 2011). Although computational modeling of related tasks (Ipeirotis et al., 2014) suggests that it should be possible for crowdsourcing to yield speech ratings of comparable accuracy or validity to those assigned by trained listeners, this question has never been investigated empirically.

The present paper investigates this question using the crowdsourcing platform that is best-developed and most widely used at the present time, Amazon Mechanical Turk (AMT). The first section provides an overview of AMT and discusses its advantages and limitations as a tool for research. Subsequent sections describe a specific mechanism developed for online collection of ratings of speech sounds and report the results of a comparison of ratings obtained from trained listeners versus non-specialists recruited through AMT. These findings suggest that crowdsourcing can represent an efficient means of obtaining valid perceptual outcome measures for speech treatment research. For readers interested in adopting AMT for their own research, the appendix offers practical guidelines for setting up data collection on AMT.

1.1. Amazon Mechanical Turk

Amazon Mechanical Turk (AMT) is a crowdsourcing platform where employers can post electronic tasks and members of the AMT community can sign up to complete tasks for pay. Previous work in other disciplines, such as cognitive science and linguistics, has offered in-depth overviews of AMT and guidelines for its use in research. This section summarizes this body of work for the audience of researchers in communication sciences and disorders, drawing in particular on Mason and Suri (2012). The original ‘‘Mechanical Turk’’ was a supposed chess-playing automaton that gained fame in the late 1700s for defeating such formidable opponents as Benjamin Franklin and Napoleon Bonaparte. In the 1820s, it was revealed that the Turk was not a true automaton but was operated by a human chess master hidden inside the machinery. Like the original Mechanical Turk, AMT is characterized as ‘‘artificial artificial intelligence’’—tasks are performed as if automated, but with human intelligence as the driving force. Amazon originally developed AMT as an internal platform for routine tasks that computers do not perform effectively, such as identifying objects in photographs. Now anyone can sign up to post tasks or complete jobs using AMT’s standardized searchable interface. Jobs on AMT are referred to as Human Intelligence Tasks (HITs). Most HITs require only seconds to minutes to complete, and compensation for these microtasks is typically only a few cents. By completing large numbers of HITs, which tend to be simple and repetitive, workers earn an average effective hourly wage that has been estimated at $4.80 (Ipeirotis, 2010a).

The AMT interface is used by hundreds of thousands of workers, along with roughly 10,000 requestors or employers (Ipeirotis, 2010a; Mason & Suri, 2012). It is possible to complete tasks on AMT from any country, but most workers come from the US and India. Employers can impose geographic restrictions on workers, e.g., allowing only workers with a US-based IP address to view a posted HIT. Based on the self-reported demographics of approximately 3000 workers across five different studies, AMT workers are more likely to be female than male, and they have a median age of roughly 30 years (Suri & Watts, 2011). Most workers do not rely on AMT as their primary source of income; in a survey, nearly 70% of US-based workers selected the statement ‘‘fruitful way to spend free time and get some cash’’ to characterize their motivation for taking jobs on AMT (Ipeirotis, 2010b).