Assessing the validity of a crowdsourcing approach to speech data rating

Tara McAllister Byun,[1] Peter Halpin,[2] and Daniel Szeredi[3]

[1] *Department of Communicative Sciences and Disorders, NYU Steinhardt School of Culture, Education, and Human Development, New York, NY 10012 USA*

[2]*Department of Humanities and Social Sciences, NYU Steinhardt School of Culture, Education, and Human Development, New York, NY 10012 USA*

[2]*Department of Linguistics, NYU College of Arts and Science, New York, NY 10012 USA*

Collecting independent listeners' judgments of the accuracy or intelligibility of speech is an essential component of research on speech disorders. However, obtaining these ratings can require considerable time and expense. The advent of crowdsourcing platforms such as Amazon Mechanical Turk (AMT) has created an unprecedented opportunity for researchers to collect data from hundreds of listeners in a rapid, inexpensive fashion. We developed an online system for sound file playback and response collection using the Experigen platform (Becker & Levine, 2010). Via AMT, we recruited 200 listeners to assign binary accuracy ratings to 100 /r/ sounds produced by children with /r/ misarticulation. Ratings were completed in 23 hours and cost $165. To analyze the rating data, we fitted listeners' binary responses as a logistic function of F3–F2 distance across /r/ tokens. We took the mode across (a) three trained raters, (b) different-sized groups of AMT raters, generated by random selection with replacement. For n(raters) > 13, the mode across AMT raters equaled or exceeded the performance of the mode across 3 experts in 80% of sampling runs. Because samples of this size are easily obtained, we advocate for further investigation and adoption of crowdsourcing approaches to speech data rating.

Word count: 199

August 28, 2013