The Wisdom of the Crowd's Ear: Speech Accent Rating and Annotation with Amazon Mechanical Turk

Stephen A. Kunath
Linguistics Department
Georgetown University
Washington, D.C. 20057
sak68@georgetown.edu

Steven H. Weinberger
Program in Linguistics
3e4 George Mason University
Fairfax, VA 22030
weinberg@gmu.edu

Abstract

Human listeners can almost instantaneously judge whether or not another speaker is part of their speech community. The basis of this judgment is the speaker's accent. Even though humans judge speech accents with ease, it has been tremendously difficult to automatically evaluate and rate accents in any consistent manner. This paper describes an experiment using the Amazon Mechanical Turk to develop an automatic speech accent rating dataset.

1 Introduction

In linguistics literature and especially in second language acquisition research, the evaluation of human speech accents relies on human judges. Whenever humans listen to the speech of others they are almost instantly able to determine whether the speaker is from the same language community. Indeed, much of the research in accent evaluation relies on native speakers to listen to samples of accented speech and rate the accent severity (Anderson-Hsieh,et. al., 1992; Cunningham-Anderson and Engstrand 1989; Gut, 2007; Koster and Koet 1993; Magen, 1998, Flege, 1995; Munro, 1995, 2001). Two problems arise from the use of this methodology. One is that the purely linguistic judgments may be infiltrated by certain biases. So for example, all other things being equal, some native English judges may interpret certain Vietnamese accents as being more severe than say, Italian accents when listening to the English uttered by speakers from these language backgrounds. The second, and more theoretically interesting problem, is that human judges make these ratings based upon some hidden, abstract knowledge of phonology. The mystery of what this knowledge is and contains is real, for as Gut (2007) remarks, "…no exact, comprehensive and universally accepted definition of foreign accent exists" (p75). The task of this linguistic and computational study is to aid in defining and uncovering this knowledge.

This study aims to develop a method for integrating accent ratings and judgments from a large number of human listeners, provided through Amazon Mechanical Turk(MTurk), to construct a set of training data for an automated speaker accent evaluation system. This data and methodology will be a resource that accent researchers can utilize. It reflects the wisdom of the crowd's ear to help determine the components of speech that different listeners use to rate the accentedness of non-native speakers.

2 Source Data

This task required HIT workers to listen to and rate a selection of non-native English speech samples. The source of all the speech samples for this effort was George Mason University's Speech Accent Archive (http://accent.gmu.edu). The Speech Accent Archive was chosen because of the high quality of samples as well as the fact that each speech

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 168–171,
Los Angeles, California, June 2010.

sample had readings of the same elicitation paragraph. This elicitation paragraph was designed to include all of the phonological features considered part of native English speech. Additionally, narrow phonetic transcriptions and phonological generalizations are available for each sample. Each speaker's information record contains demographic information and language background information. Three native language groups were selected for this study: Arabic, Mandarin, and Russian. The motivation for this particular selection comes from the fact that each of these languages represents a different language family. These languages contain different phonetic inventories as well as phonological patterns.

3 HIT Description

Our HIT consisted of three sections. The first section asked the worker to describe their own native language background and any foreign language knowledge or experience. Asking about native and foreign language experience allowed us to estimate possible rating bias arising from experience with second language phonology. The second section of the HIT included two rating tasks for use as a baseline and to help the workers get acclimated to the task. Each worker was asked to listen to two audio samples of speakers reading the same elicitation paragraph, one of a native English speaker and one of a native Spanish speaker who started learning English late in life. The rating scale used was a five point Likert scale. After completing the baseline question, workers began the third section and were then asked to listen to fifteen samples of non-native English speakers read the same elicitation paragraph. After listening to each sample the workers were asked to rate the accentedness of the speech on the five point Likert scale. The five-point scale rates native accent as a 1 and heavy accent as a 5. Workers were additionally asked to group each speech sample into different native language categories. For this question they were presented with 3 language family groups: A, B, and C. Based on their perception of each speech sample they would attempt to categorize the fifteen speakers into distinct groups native language groups.

4 Worker Requirements and Cost

Due to the type of questions contained in our HIT we came up with several worker requirements for the HIT. The first and most important requirement was that HIT workers be located inside of the USA so as to limit the number of non-native English speakers. This requirement also helped to increase the likelihood that the listener would be familiar with varieties of English speech accents common in America. Additionally, due to the size of the task we had a requirement that any worker must have at least a 65% approval record for previous HITs on other MTurk tasks. After looking at other-comparably difficult tasks we decided to offer our first HIT at $0.75. Subsequent HITs decreased the offered price to $0.50 for the task.

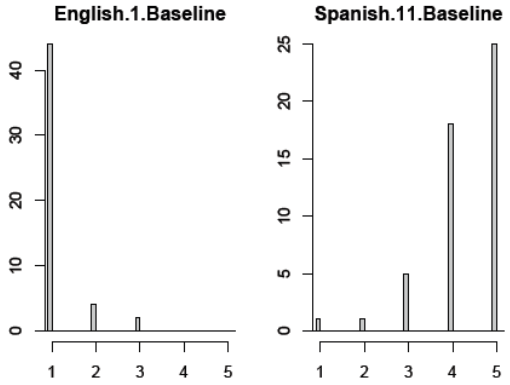

Figure 1. Mechanical Turk workers ratings of 2 baseline samples: English 1 and Spanish 11. The numbers on the horizontal axis represent the how native-like the speaker was rated. A (1) indicates that the speaker sounds like a native English speaker. A (5) indicates the presence of a heavy accent.

5 HIT Results

Two HITs were issued for this task. Each HIT had 25 workers. Average time for each worker on this task was approximately 12.5 minutes. Initial data analysis showed that users correctly carried out the tasks. Baseline question results, shown in figure 1, indicated that virtually every worker agreed that the native English speaker sample was a native speaker of English. The ratings of the baseline Spanish showed that workers generally agreed that it was heavily accented speech. In addition to the

high quality of baseline evaluations, workers consistently provided their own native and foreign language information.

Ratings of the speech samples in each question, as seen in Figure 2, showed relatively consistent evaluations across workers. A more detailed statistical analysis of inter-worker ratings and groupings is currently underway, but the initial statistical tests show that there was a consistent correlation between certain phonological speech patterns and ratings of accentedness.

6 Future Work

This experiment has already provided a wealth of information on how human's rate accents and how consistent those ratings are across a large number of listeners. Currently, we are integrating the accent ratings with the phonetic transcriptions and the list of identified phonological speech processes to construct a set of features that are correlated with accent ratings. We have begun to capitalize on the Mechanical Turk paradigm and are constructing a qualification test to help us better understand inter-worker agreement on accent rating.

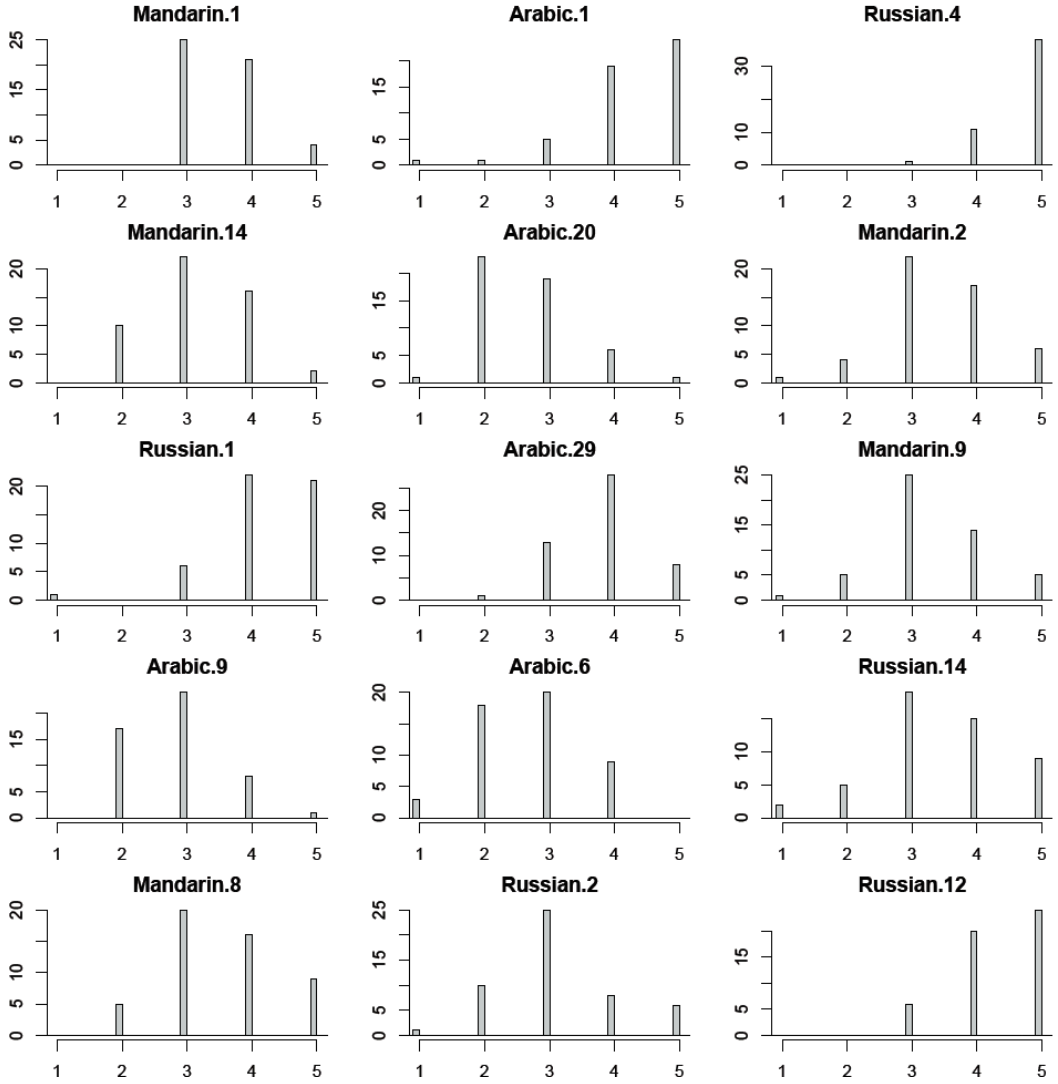

Figure 2. Workers accent ratings for all speech samples. The horizontal axis indicates the accentedness rating: (1) is a native English accent and (5) is heavily accented. The vertical axis indicates the number of HIT workers that provided the same rating for the sample. The numbers at the end of each language name represent the Speech Accent Archive sample id for the language, e.g. Mandarin.1 indicates that the sample was the Mandarin 1 speaker on the Archive.

This qualification test will include a larger sample of Native English speech data as well as a broader selection of foreign accents. In this new qualification test workers will be presented with a scale to rate the speakers accent from native-like to heavily accented. Additionally, the user will be asked to group the samples into native language families. Once the user passes this qualification test they will then be able to work on HITs that are considerably shorter than the original long-form HIT described in this paper. In the new HITs workers will listen to one or more speech samples at a time and both rate and, if required, attempt to group the sample relative to other speech samples. The selection criteria for these new samples will be based on the presence of phonological speech processes that have the highest correlation with accent ratings.

Acknowledgments

The authors would like to thank Amazon.com and the workshop organizers for providing MTurk credits to perform this research.

References

Anderson-Hsieh, J., Johnson, R., & Kohler, K. 1992. The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning,* 42, 529-555.

Cunningham-Anderson, U., and Engstrand, E., 1989. Perceived strength and identity of foreign accent in Swedish. *Phonetica*, 46, 138-154.

Flege, James E., Murray J. Munro, and Ian R.A. MacKay (1995). Factors Affecting Strength of Perceived Foreign Accent in a Second Language. *Journal of the Acoustical Society of America*, 97, 5, pp 3125-3134.

Gut, U. 2007. Foreign Accent. In C. Muller, (ed.), *Speaker Classification I*. Berlin: Springer.

Koster, C., and Koet, T. 1993. The evaluation of accent in the English of Dutchmen. *Language Learning*, 43, 1, 69-92.

Lippi-Green, R. 1997. *English with an Accent*. New York: Routledge.

Magen, H. 1998. The perception of foreign-accented speech. *Journal of Phonetics*, 26, 381-400.

Munro, Murray J. (1995). Nonsegmental Factors in Foreign Accent: Ratings of Filtered Speech. *Studies in Second Language Acquisition,* 17, pp 17-34.

Munro, Murray J. and Tracey M. Derwing (2001). Modeling Perceptions of the Accentedness and Comprehensibility of L2 Speech: The Role of Speaking Rate. *Studies in Second Language Acquisition,* 23, pp 451-468.

Scovel, T. 1995. Differentiation, recognition, and identification in the discrimination of foreign accents. In J.Archibald (ed), *Phonological Acquisition and Phonological Theory*. Hillsdale, NJ: Lawrence Erlbaum.