REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
25-10-2010	Technical Paper	OCT 2010 - NOV 2010

4. TITLE AND SUBTITLE

Assessing The Speaker Recognition Performance Of Naive Listeners Using Mechanical Turk

5a. CONTRACT NUMBER
FA8720-05-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wade Shen, Joseph Campbell, Derek Straub, and Reva Schwartz

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

NSA
9800 Savage Rd
Ft. Meade, MD 20755

10. SPONSOR/MONITOR'S ACRONYM(S)
NSA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this paper we attempt to quantify the ability of naive listeners to perform speaker recognition in the context of the NIST evaluation task. We describe our protocol: a series of listening experiments using large numbers of naive listeners (432) on Amazon's Mechanical Turk that attempt to measure the ability of the average human listener to performance speaker recognition. Our goal was the compare the performance of the average human listener to both forensic experts and state-of-the-art automatic systems. We show that naive listeners vary substantially in their performance, but that a voting of listeners can achieve performance similar to that of expert forensic examiners.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Zach Sweet
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	SAR	4	**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U			781-981-5997

THIS MATERIAL HAS BEEN CLEARED
FOR PUBLIC RELEASE BY 66 ABW/PA
DATE: 25Oct10
CASE # 66ABW-2010-1258

MS-51985

ASSESSING THE SPEAKER RECOGNITION PERFORMANCE OF NAIVE LISTENERS USING MECHANICAL TURK

Wade Shen, Joseph Campbell, Derek Straub

MIT/Lincoln Laboratory
244 Wood Street
Lexington, MA 02476
USA

Reva Schwartz

United States Secret Service
2255 H St. NW
Washington, DC 20006
USA

ABSTRACT

In this paper we attempt to quantify the ability of naive listeners to perform speaker recognition in the context of the NIST evaluation task. We describe our protocol: a series of listening experiments using large numbers of naive listeners (432) on Amazon's Mechanical Turk that attempt to measure the ability of the *average* human listener to performance speaker recognition. Our goal was the compare the performance of the average human listener to both forensic experts and state-of-the-art automatic systems. We show that naive listeners vary substantially in their performance, but that a voting of listeners can achieve performance similar to that of expert forensic examiners.

Index Terms— One, two, three, four, five

1. INTRODUCTION

It is commonly hypothesized that the sound of the human voice is a characteristic of a it's speaker's identity and that these characteristics are perceivable by human listeners. Research into automatic methods have systematically shown that acoustic features, phonetic and word usage can all yield varying degrees of speaker identifiability [1, 2], but comparatively few studies have been conducted to assess the ability of human listeners to identify speakers on a large scale [3].

Despite the lack of empirical evidence, this hypothesis has been widely accepted as fact in the forensic community. It has given rise to the to the discipline of forensic speaker recognition as conducted by human experts, in which audio samples from known and unknown sources are compared by an expert through the process of listening. In this community it is often assumed that the ability to listen for speaker identity requires training and the application specialized identification processes, though little scientific validation has been done to prove the efficacy of human perception used in many of these methods [4, 5]. In fact, many more studies have limitations of human perceptions especially, when voice samples are short, stressed, unfamiliar, disguised or noise-corrupted [6, 7, 8, 9].

As part of the NIST 2010 Speaker Recognition evaluation, NIST conducted a first-of-its-kind systematic benchmark test to assess the ability of human listeners and machine algorithms to perform speaker recognition.

In this paper we attempt to quantify the ability of naive listeners to perform speaker recognition in the context of the NIST evaluation task. We describe our protocol: a series of listening experiments using naive listeners on Amazon's Mechanical Turk that attempt to measure the ability of the *average* human listener to performance speaker recognition. Our goal was the compare the performance of the average human listener to both forensic experts and state-of-the-art automatic systems. This section describes the experiments that we ran and some preliminary conclusions based on the results we obtained.

The experiments conducted as part of this work focused on the following main areas:

- **Elicitation:** How to structure the listening task in a way that subjects perform to their optimal abilities.
- **Scoring:** How to assign scores to speaker verification trials and aggregate those scores across subjects
- **Preliminary Measurement:** Once the above issues were addressed, we ran an experiment to quantify human performance

2. MECHANICAL TURK

Amazon's Mechanical Turk system provides a mechanism for payment and recruiting of human labor for tasks that can be conducted online. The system allows *requesters* to create labeling/annotation tasks, forms, surveys, etc. that can be distributed to a large pool of workers in the US and around the world. Tasks may be arbitrarily small in terms of required

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

effort from workers (e.g. image labeling) and the system handles accounting and payment for potentially large sets of tasks. This allows for researchers to conduct many trials without significant bookkeeping.

Mechanical Turk is a market-driven system: potential subjects (called workers) can see tasks descriptions and payment information before choosing what to work on. As there are often tens of thousands of workers available at any given time, the cost of annotation can be very low and the turnaround time for conducting experiments can be very fast. Research collaborators at MIT/BCS have averaged $0.87 per hour from psycho-linguistic experiments they have been conducting over the past two years. This is significantly lower than the cost of running live human subjects in the lab.

2.1. Turk-specific Issues

Despite the ease-of-use and lowered subject costs, Mechanical Turk does offer less controls than human subject experiments run in the lab. Many experiments have observed that motivation and accuracy issues are prevalent.

Since tasks compete with each other for workers, proper pricing is important in order to ensure that subjects perform your task accurately and quickly.

Because tasks are often priced in terms of the number of completed tasks/annotations/surveys, subjects are often motivated to finish these tasks as quickly as possible (to maximize their effective hourly rate).

Proper task design for Mechanical Turk is required to ensure that subjects are willing to work on your task and that they complete your task as accurately as possible. The later is especially difficult to enforce without some mechanism to verify task results.

Mechanical Turk offers very little in terms of subject biographical data. As a result it is difficult to control for gender/age and other external factors. For our particular experiment, we would prefer that subjects be native American English speakers so as to eliminate potential cross-language speaker-verification performance issues. Mechanical Turk provides information about whether a worker is located within the US and it allows us to filter workers on this basis. Any further biographical information regarding nativeness would need to be collected during the experiment and trials for non-natives would require filtering after payment.

3. EXPERIMENT SETUP

Since our goal was to compare the results of our Mechanical Turk experiments with the HASR submissions to the NIST 2010 SRE, we adapted NIST's verification protocol. In order to reach the maximal number of subjects, each NIST verification trial was presented as a separate task to Mechanical Turk workers. Potential workers could do each trial exactly once, but were not required to do all trials. For each trial, subjects were asked to listen to the two NIST-supplied audio clips. As preprocessing, the speaker-of-interest was extracted and all audio levels were normalized to -8db. For each trial, we maintain results per subject and the amount of time each subject required to complete the trial.

In order to motivate subjects to listen to both audio clips in their entirety, included a set of listening comprehension questioned asking about facts that are stated at 1-minute intervals in the audio. The inclusion of these questions increased the average amount a subject spent per trial from less than 2 minutes to 7 minutes. Furthermore, subjects were asked to provide qualitative confidence assessments about each trial. We conducted experiments using two different scales:

1. A Likert-like scale (1-5) as shown in figure 2.
2. We asked subjects to assign a % confidence that the "Two audio clips were from the same speaker."
3. A hard decision with an additional confidence scale (3-point, see figure 1).

We asked subjects to be as accurate as possible in both their listening comprehension questions and their trial decisions and we conditioned their payment on accuracy. Each trial was priced at $0.33 with an effective hourly rate of $2.82/hour. Guidelines for the experiment are shown in figure 1. Figure 2 shows the display for a given trial with the Likert-like scale used for that set of experiments.

As subjects may have scale biases/ranges, we encouraged subjects to complete all 15 trials and were paid a bonus ($1.00) to do so.

4. RESULTS

We ran three sets of experiments using the the scales reported above (150 trials for scales 1 and 2, 300 trials for scale 3). In total more than 600 trials were conducted using 432 Mechanical Turk subjects. We assessed the performance of the average human by weighted voting: scores from every subject were first normalized to zero-mean/unit-variance. Then the resulting z-scores per trial were averaged. For each scoring variant a threshold was set to minimize the total cost (C_{total}) where: $C_{total} = N_{fa} + N_{miss}$ (This scoring assumes equal cost of miss and false alarm).

Table 1 shows the results for the different scales. Interestingly, the Mechanical Turk listeners were very close, in performance, to the average amongst HASR participants (Mechanical Turk: 6-7 errors, HASR Average: 6.6 errors per submitted system). These results may improve with more normalization data per subject (1.38 trials per subject on average). That said, these numbers are quite a bit worse than our best automatic systems (which make only one error on this set). Because of the peculiar way in which these trials were selected, we hope to run a follow on experiment using more

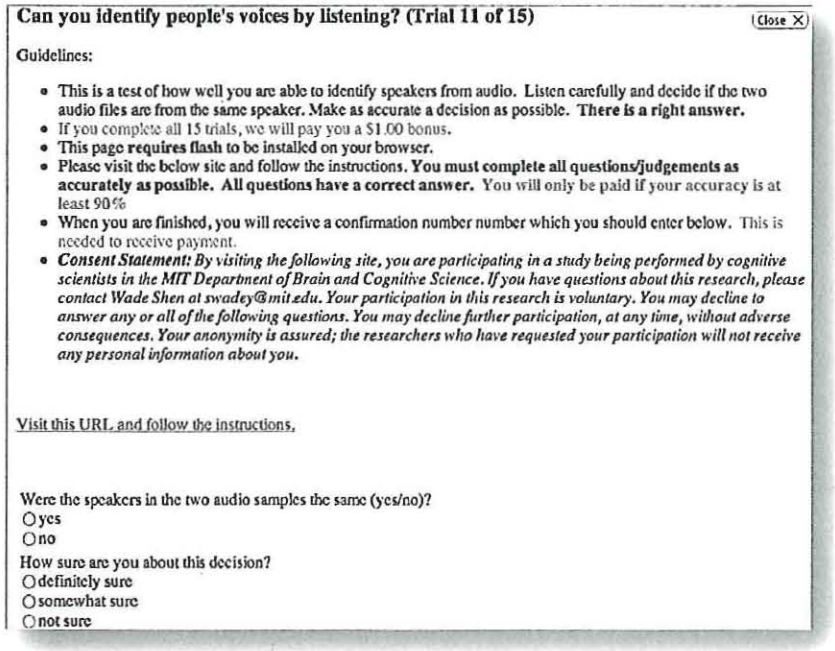

Fig. 1. Instructions and guidelines presented to subjects

Scale for Subjects	Optimal FAs	Optimal Misses
Confidence Only	1	5
Likert Scale	3	4
Hard Decision + Confidence	1	5

Table 1. Comparison of scoring scale performance for Mechanical Turk trials

typical trials. We expect that the gap between average listeners and machines will narrow.

Relatively few subjects (28 in total) completed all trials. More subjects and more trial pairs are needed (to be collected in future experiments) to do a reliable analysis of individual subject performance. That said, the data in table 2 suggests that our within-subject normalization scheme may be effective. Voting using normalized scores across all trials from subjects appears to improve the performance over that of the average subect and is close in performance to the best subject.

5. DISCUSSION

From the limited trial set, we learned that naive listeners (especially panels of such listeners with proper normalization) can perform speaker recognition on par with forensic experts. In results reported by NIST, sites using human listeners exclusively exhibit similar C_{total} numbers. Interestingly, the best automatic systems make relatively fewer errors (C_{total} = 1 for MIT/LL's best system). This may be due to compensation methods developed for cross-channel trials which are prevalent in this data set. The selected trial set was choosen to be exceptionally difficult for human listeners. Given the small data set, it's not clear that these trials are particularly difficult for automatic methods. A more randomly selected trial set is needed to assess if these data are equally difficult for both methods.

Because our method for focusing subjects required listening comprehension questions written from transcripts, we were limited to the small subset of HASR1 trials. In future experiments, we would like to expand the trial set for more statistical reliability.

Our protocol also makes no attempt to find "good" or "trained" human listeners. In future experiments it would be possible to find a subset of listeners that meet a specific performance criteria on non-HASR data and assess their performance on this task. This adjusted protocol could be used to assess limits of human performance.

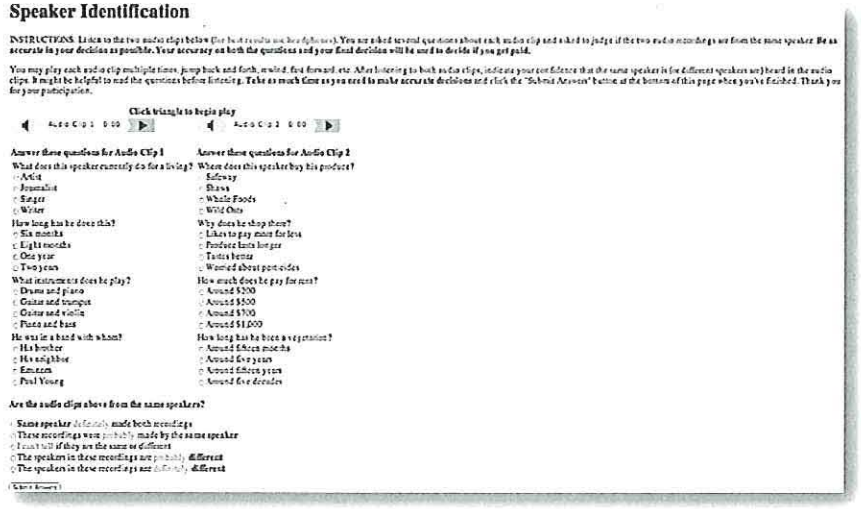

Fig. 2. Display of what a typical Mechanical Turk trial looked like for subjects

		FAs	Misses
Individuals	Worst Subject	5	4
	Best Subject	1	4
	Average Subject	2.8	4.4
Voted (Optimal FA/Miss)	–	1	5

Table 2. Comparison of individual subject performance vs. voted average

6. REFERENCES

[1] Douglas Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andre Adami, Qin Jin, David Klusacek, Joy Abramson, Radu Mihaescu, Jack Godfrey, Doug Jones, and Bing Xiang, "The supersid project: Exploiting high-level information for high-accuracy," in *In Proc. International Conference on Audio, Speech, and Signal Processing, Hong Kong*, 2003, pp. 784–787.

[2] Mark Przybocki Alvin and Alvin Martin, "Nist speaker recognition evaluation chronicles," in *Proc. Odyssey 2004, The Speaker and Language Recognition Workshop*, 2004, pp. 12–22.

[3] H. Hollien and R. Schwartz, "Speaker identification utilizing noncontemporary speech," *Journal of Forensic Science*, vol. 46, pp. 63–67, 2001.

[4] Peter Ladefoged and J. Ladefoged, "The ability of listeners to identify voices," *UCLA Working Papers in Phonetics*, 1980.

[5] Astrid Schmidt-nielsen and Thomas H. Crystal, "Human vs. machine speaker identification with telephone speech," *Digital Signal Processing*, vol. 10, 1998.

[6] A. Compton, "Effects of filtering and vocal duration upon the identification of speakers, aurally," *Journal of the Acoustical Society of America*, vol. 35, pp. 1748–1752, 1963.

[7] Daniel Yarmey, "Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations," *International Journal of Speech Language and the Law*, vol. 11, no. 2, 2004.

[8] A.R. Reich and J.E. Duke, "Effects of selective vocal disguise upon speaker identification by listening," *Journal of the Acoustical Society of America*, vol. 66, 1979.

[9] H. Hollien, W. Majewski, and E.T. Doherty, "Perceptual identification of voices under normal, stress and disguise speaking conditions," *Journal of Phonetics*, vol. 10, 1982.