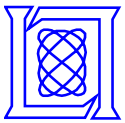


USSS-MITLL 2010 Human-Assisted Speaker Recognition: Mechanical Turk (Part 2)

Wade Shen, Reva Schwartz (USSS), Derek Straub, Joseph Campbell

jpc@ll.mit.edu

NIST Speaker Recognition Evaluation Workshop
Brno, Czech Republic

25 June 2010

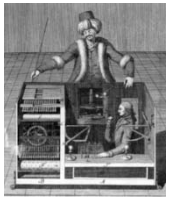
This work was sponsored by the Department of Defense under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Outline

- Introduction
- Audio Preprocessing
- HASR System
- Qualitative
- Automatic
- Fusion
- Results
- **Mechanical Turk**
- **Summary**

Mechanical Turk HASR

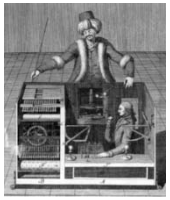
- Give HASR1 to many human subjects
  - Measure naïve listener performance on speaker comparison
  - Amazon's Mechanical Turk enables this*
    - IRB compliant
- Focus on experimental design
  - Motivate subjects
  - Careful subject interaction
- Design scales tried thus far
  - **Confidence question:** subjects assign 0% to 100% confidence per trial (Likert-like values: 0%, 25%, 50%, 75%, 100%)
  - **Hard decision + confidence:** subjects make hard decision and assign confidence %
  - **Standard Likert item scale:** 1. strongly disagree, 2. disagree, 3. neither agree nor disagree, 4. agree, 5. strongly agree



*Thanks to Ted Gibson, MIT, for sharing his account. https://www.mturk.com/mturk/

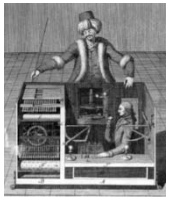
Turk-Specific Issues

- Force subjects to listen to the audio
  - Reading comprehension questions
  - Questions test different portions of audio file
  - 8-10 questions per trial
  - Average time per trial:
    - 7 min (w/questions), < 2 min (w/o) questions
- Reward performance
  - Subjects only paid if comprehension > 90%
  - Bonus payment for completing all 15 trials
- Present and protect audio samples
  - Flash audio player interface: intuitive and hard to rip
    - No requests for waveform or spectrogram display
  - Headphones encouraged
  - Audio samples are our preprocessed HASR1 samples: mono, purified, enhanced (interview) or echo canceled (telephone)



Subjects

- Subjects from US (IP address filter), but could be non-native
  - Subjects must pass the listening comprehension questions
  - English listening skills must be good to handle CTS
- Pay: $0.33/trial + $1.00 bonus for full set of 15 trials
  - A relatively high-paying, but high-effort, MT task
- Repeat subjects across each scoring method not counted beyond first method
  - But these subjects may have had more listening time: reanalysis needed



What Subjects See

Speaker Identification

Listening Page

INSTRUCTIONS: Listen to the two audio clips below (for best results use headphones). You are asked several questions about each audio clip and asked to judge if the two audio recordings are from the same speaker. **Be as accurate in your decision as possible. Your accuracy on both the questions and your final decision will be used to decide if you get paid.**

You may play each audio clip multiple times, jump back and forth, rewind, fast-forward, etc. After listening to both audio clips, indicate your confidence that the same speaker is (or different speakers are) heard in the audio clips. It might be helpful to read the questions before listening. **Take as much time as you need to make accurate decisions** and click the "Submit Answers" button at the bottom of this page when you've finished. Thank you for your participation.

Click triangle to begin play

Audio Clip 1 0:00 Audio Clip 2 0:00

Answer these questions for Audio Clip 1

What does this speaker currently do for a living?
- Artist
- Journalist
- Singer
- Writer

How long has he done this?
- Six months
- Eight months
- One year
- Two years

What instruments does he play?
- Drums and piano
- Guitar and trumpet
- Guitar and violin
- Piano and bass

He was in a band with whom?
- His brother
- His neighbor
- Eminem
- Paul Young

Answer these questions for Audio Clip 2

Where does this speaker buy his produce?
- Safeway
- Shaws
- Whole Foods
- Wild Oats

Why does he shop there?
- Likes to pay more for less
- Produce lasts longer
- Tastes better
- Worried about pesticides

How much does he pay for rent?
- Around $200
- Around $500
- Around $700
- Around $1,000

How long has he been a vegetarian?
- Around fifteen months
- Around five years
- Around fifteen years
- Around five decades

Are the audio clips above from the same speakers?

- **Same speaker** definitely made both recordings
- These recordings were probably made by the **same speaker**
- I can't tell if they are the same or different
- The speakers in these recordings are probably **different**
- The speakers in these recordings are definitely **different**

Submit Answers

Can you identify people's voices by listening? (Trial 11 of 15)

Close

Guidelines:

- This is a test of how well you are able to identify speakers from audio. Listen carefully and decide if the two audio files are from the same speaker. Make as accurate a decision as possible. **There is a right answer.**
- If you complete all 15 trials, we will pay you a $1.00 bonus.
- This page **requires flash** to be installed on your browser.
- Please visit the below site and follow the instructions. **You must complete all questions/judgements as accurately as possible. All questions have a correct answer. You will only be paid if your accuracy is at least 90%**
- When you are finished, you will receive a confirmation number number which you should enter below. This is needed to receive payment.
- ***Consent Statement:*** *By visiting the following site, you are participating in a study being performed by cognitive scientists in the MIT Department of Brain and Cognitive Science. If you have questions about this research, please contact Wade Shen at swadey@mit.edu. Your participation in this research is voluntary. You may decline to answer any or all of the following questions. You may decline further participation, at any time, without adverse consequences. Your anonymity is assured; the researchers who have requested your participation will not receive any personal information about you.*

Visit this URL and follow the instructions.

Were the speakers in the two audio samples the same (yes/no)?
- yes
- no

How sure are you about this decision?
- definitely sure
- somewhat sure
- not sure

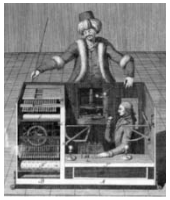
Instruction Page

Results So Far

- **Run 600 trials, 400+ subjects**
  - **Most subjects do only 1 trial (600 << 400 * 15 HASR1 trials)**
  - **Average number of trials per subject ~ 1.5**
  - **28 subjects completed all 15 trials of HASR1 (bonus increased)**
- **Results determined by weighted voting**
  - **Average subject score per trial (normalized per subject)**
  - **Optimal FA/Miss: minimize cost when**

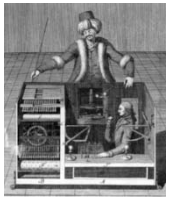
    cost(FA) == cost(Miss)
- **Results**

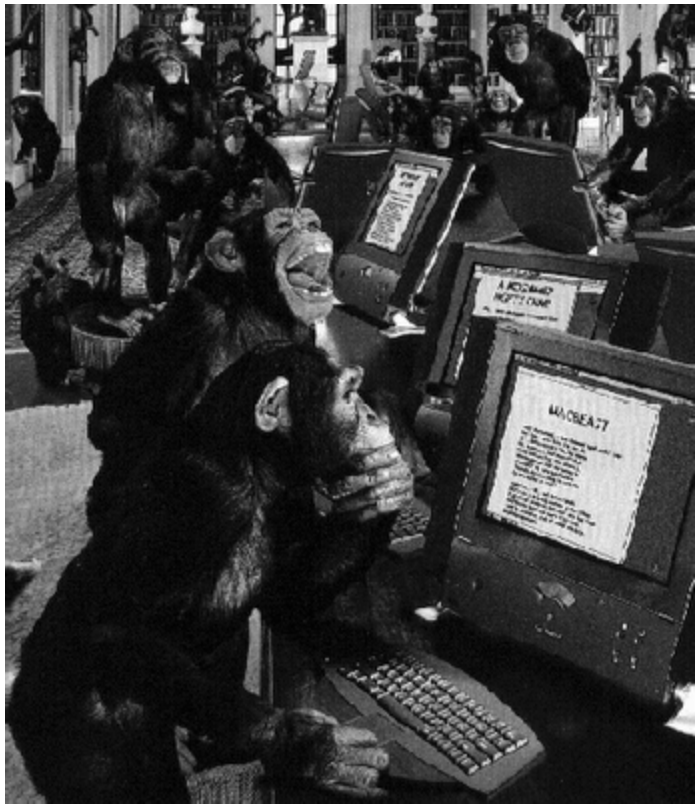
Scale for Subjects	Optimal FAs	Optimal Misses
Confidence Only	1	5
Standard Likert	3	4
Hard Decision + Confidence	1	5



Discussion

- MT results in middle of HASR pack
  - But with only naïve listeners averaging 7 minutes per trial!
  - More trials per subject for normalization (in process)
- HASR data is particularly difficult
  - What are its biases?
  - Try more typical samples?
  - Try wideband data?
  - Try alternate interview mics?
- Other challenges
  - Try fear-stressed speech?
  - Try deceptive speech?
- Line-up style experiment
  - Let subjects listen to battery of cohorts + true trial
  - Measure closed-set ID performance





Shakespeare?