MAKING THE BEST OF IMPERFECT AUTOMATIC SPEECH RECOGNITION FOR CAPTIONING ONE-ON-ONE MEETINGS

Larwan Berke, Sushant Kafle, Christopher Caulfield, Dr. Matt Huenerfauth, and Dr. Michael Stinson

NTID Scholarship Symposium - January 12, 2017

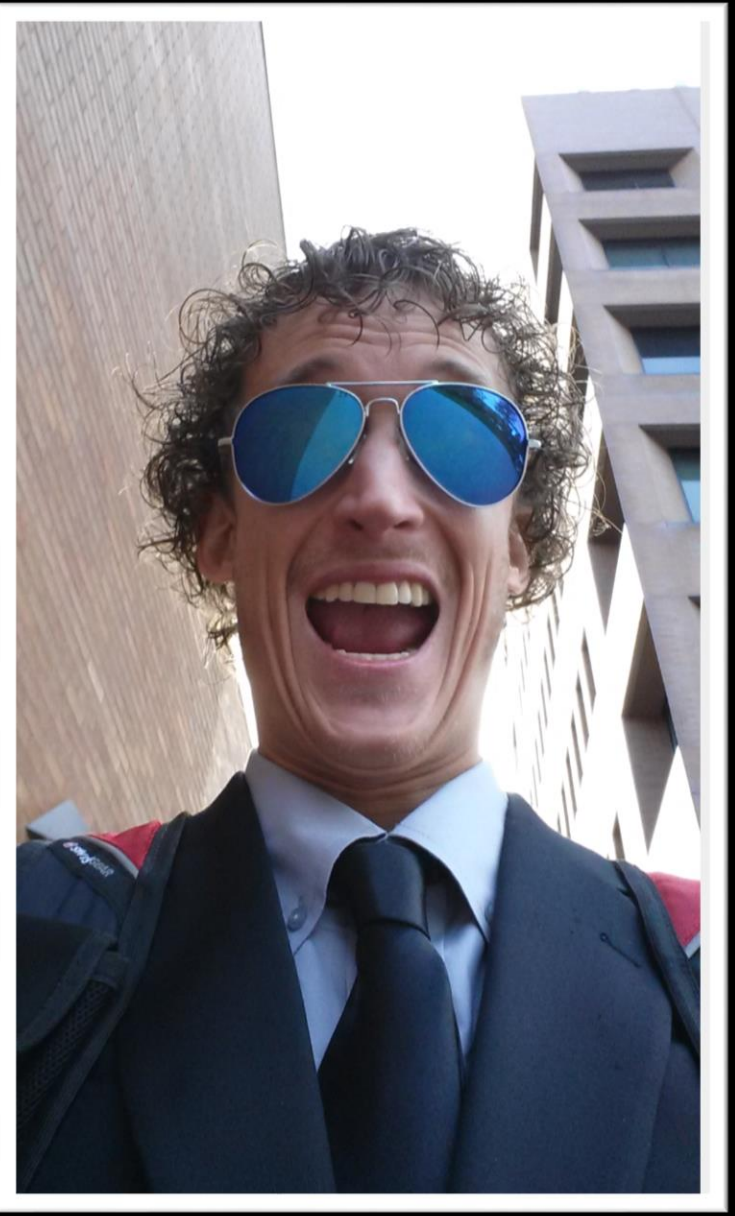
LARWAN BERKE

PH.D. STUDENT IN COMPUTING AND INFORMATION SCIENCES
ADVISOR: DR. MATT HUENERFAUTH @ GCCIS

CAIR
Center for Accessibility and Inclusion Research

The LATLab
Linguistic and Assistive Technologies Laboratory

- **Researching**: Displaying Uncertainty From Imperfect Automatic Speech Recognition For Captioning.
- **Bio**: Born Deaf to Deaf parents and grew up in Fremont, CA. Graduated with a B.S. in Mathematics from Gallaudet University in Washington, DC.
- **Why RIT**: The breadth of accessibility-related projects at RIT drew me here. It's a wonderful experience working with world-class faculty who understands the needs of the Deaf.



THE RESEARCH TEAM

Sushant Kafle
Ph.D. Student

Christopher Caulfield
Undergraduate Student

Dr. Matt Huenerfauth
Associate Professor - GCCIS

Dr. Michael Stinson
Research Faculty - MSSE

INTRODUCTION

- Around 30 million Deaf and Hard-of-Hearing (DHH) individuals (such as me!) in the US have difficulty communicating via aural means [Karchmer *et al.*, Lin *et al.*]
- Alternative means of communication: American Sign Language (ASL), Video Relay Service, E-Mail, interpretation, etc.
- High cost and limited availability of interpreters for DHH participants motivate text-based alternatives
- Some users prefer text to sign language: for example, people who become DHH later in life are less likely to use ASL

HIGH COSTS OF CAPTIONING

- Real-time stenographer impractical/too costly for many situations
- Manually captioning videos (high-quality, not real-time) typically takes 8-10 times the length of the video
- Around $1 per minute to caption videos via online services (Rev, CaptionsLab, etc.)

http://educationcareerarticles.com/wp-content/uploads/2013/02/Court-Reporter.jpg

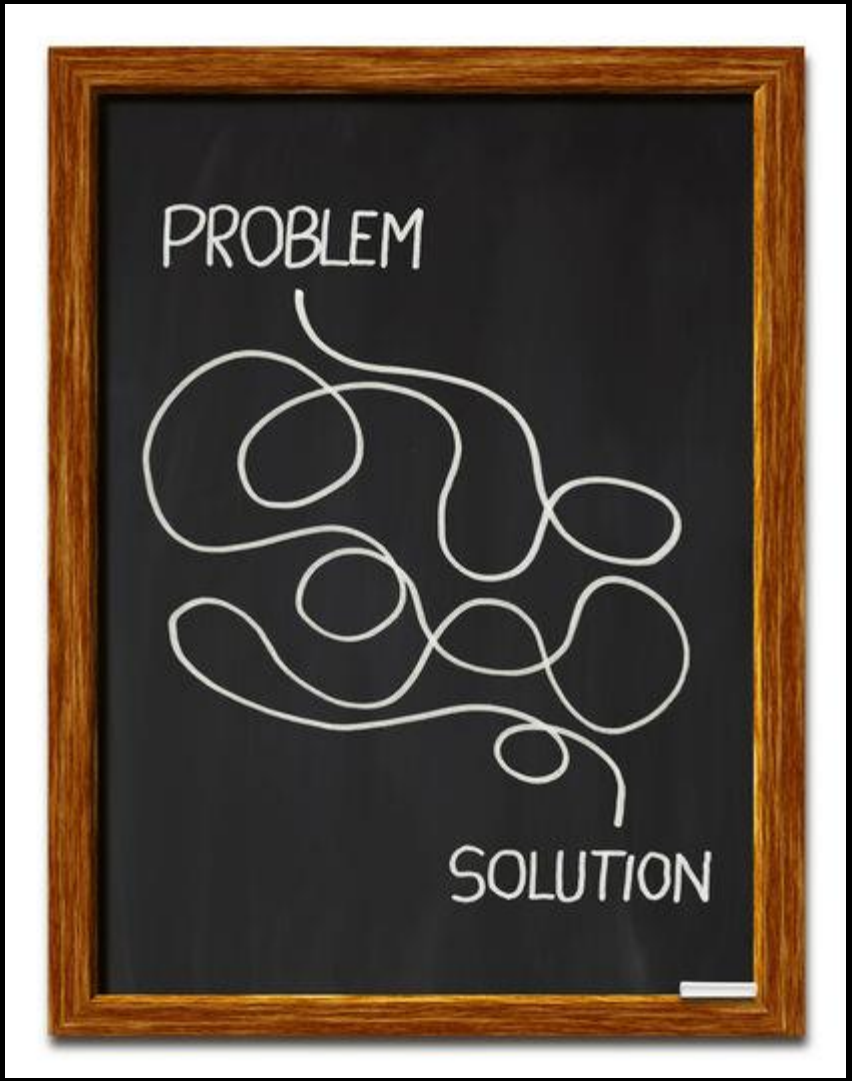
WHAT ABOUT C-PRINT?

- "C-Print is a speech-to-text (captioning) technology and service developed at the National Technical Institute for the Deaf"

- Many students are currently using the system and doing great in their classes! [Stinson, M. *et al.*]

- However, limited availability of trained captioners hinders greater adoption of this technology

https://www.rit.edu/ntid/cprint/sites/rit.edu.ntid.cprint/files//slideshow/01-Francis_Devices_Jan_13_sm_48447.jpg



HOW CAN WE SOLVE THIS PROBLEM OF AVAILABILITY AND COST?

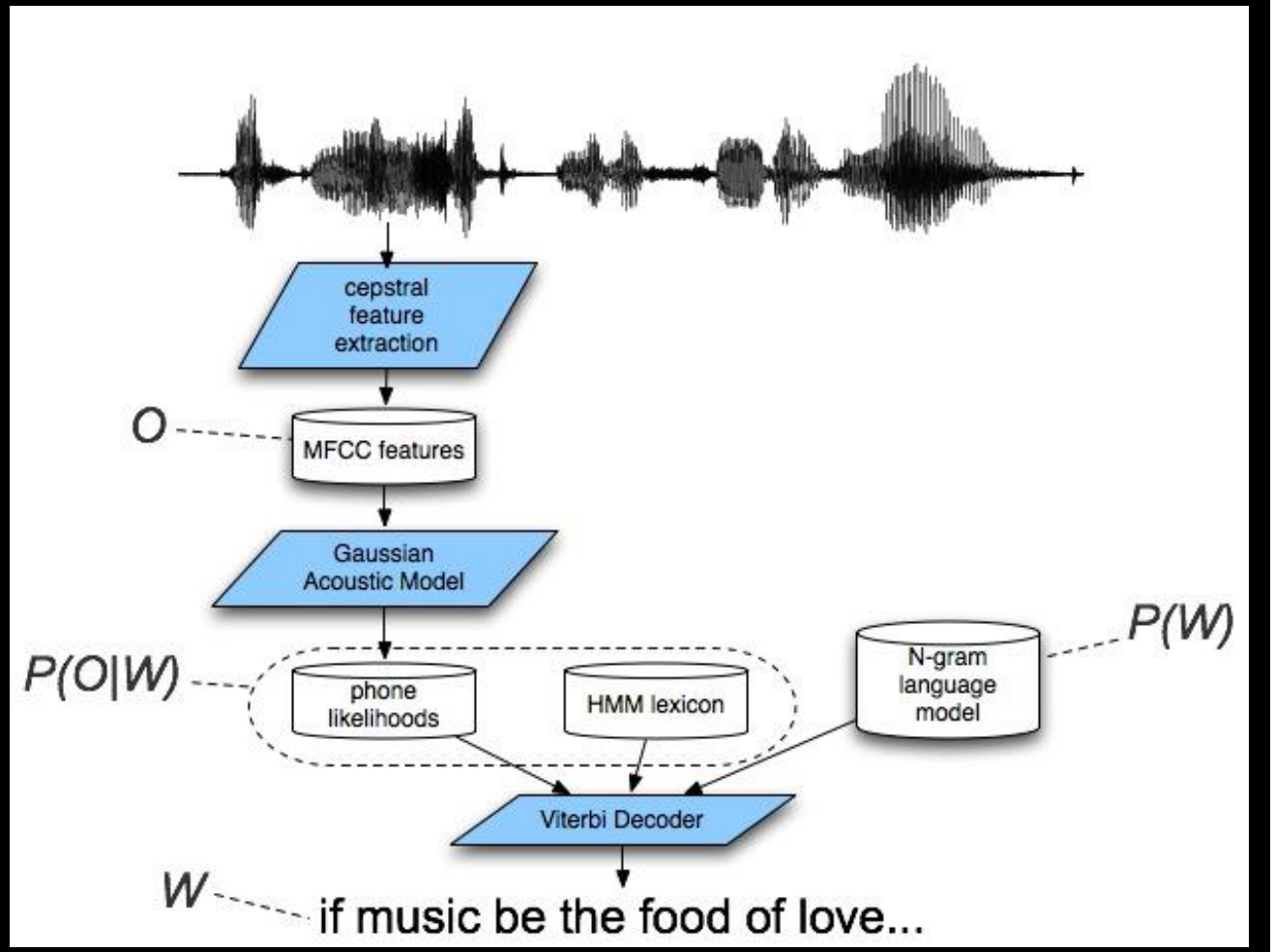
http://www.referenceforbusiness.com/photos/problem-solving-styles-174.jpg

WHAT IS AUTOMATIC SPEECH RECOGNITION?

http://www.smallbusinessbonfire.com/wp-content/uploads/2014/11/speech-recognition.jpg

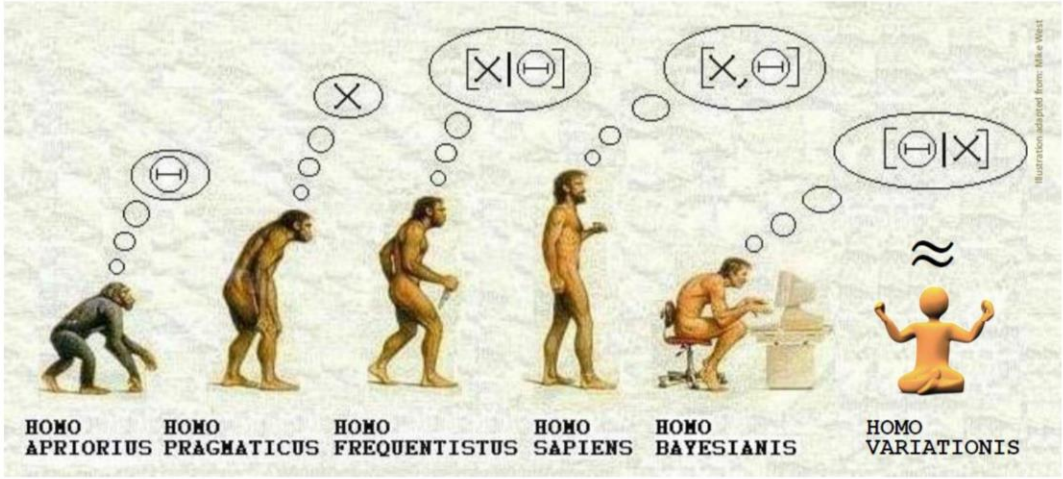
http://www.techaudible.org/wp-content/uploads/2014/05/voice-recognition.jpg

SPEECH RECOGNITION ARCHITECTURE



Third Generation ML

- Deep integration of domain knowledge and statistical learning
  - Bayesian framework
  - Probabilistic graphical models
  - Efficient inference using local message-passing



WHAT KIND OF MATH DOES ASR USE?

The first term:

$$\int_X P(X,Y,\theta^{old}) \sum_{t=1}^{T} \log P(x_t \mid x_{t-1};\theta_{x_t,x_{t-1}}) dX$$

$$= \sum_i \sum_j \sum_{t=1}^{T} P(x_t = i, x_{t-1} = j, Y, \theta^{old}) \log P(x_t = i \mid x_{t-1} = j;\theta_{i,j})$$

use a Lagrange multiplier

$$\frac{d}{d\theta_{i,j}} [\sum_i \sum_j \sum_{t=1}^{T} P(x_t = i, x_{t-1} = j, Y, \theta^{old}) \log P(x_t = i \mid x_{t-1} = j;\theta_{i,j}) - \lambda(\sum_i \theta_{ij} - 1)] = 0$$

$$\lambda = \frac{\sum_{t=1}^{T} P(x_t = i, x_{t-1} = j, Y, \theta^{old})}{\sum_{t=1}^{T} \log P(x_t = i \mid x_{t-1} = j;\theta_{i,j})}$$

$$\sum_i \theta_{i,j} = \sum_i \frac{\sum_{t=1}^{T} P(x_t = i, x_{t-1} = j, Y, \theta^{old})}{\lambda} = 1$$

$$\theta_{i,j} = \frac{\sum_{t=1}^{T} P(x_t = i, x_{t-1} = j, Y, \theta^{old})}{\sum_{t=1}^{T} P(x_{t-1} = j, Y, \theta^{old})}$$

We will discuss this part ->

WHAT KIND OF MATH DOES ASR USE?

The first term:

$$\int_X P(X,Y,\theta^{old}) \sum_{t=1}^{T} \log P(x_t \mid x_{t-1};\theta_{x_t,x_{t-1}}) dX$$

use a L

$$\frac{d}{d\theta_{i,j}}[$$

$$\lambda = -$$

$$;\theta_{i,j})$$

$$-1)] = 0$$

$$,\theta^{old})$$

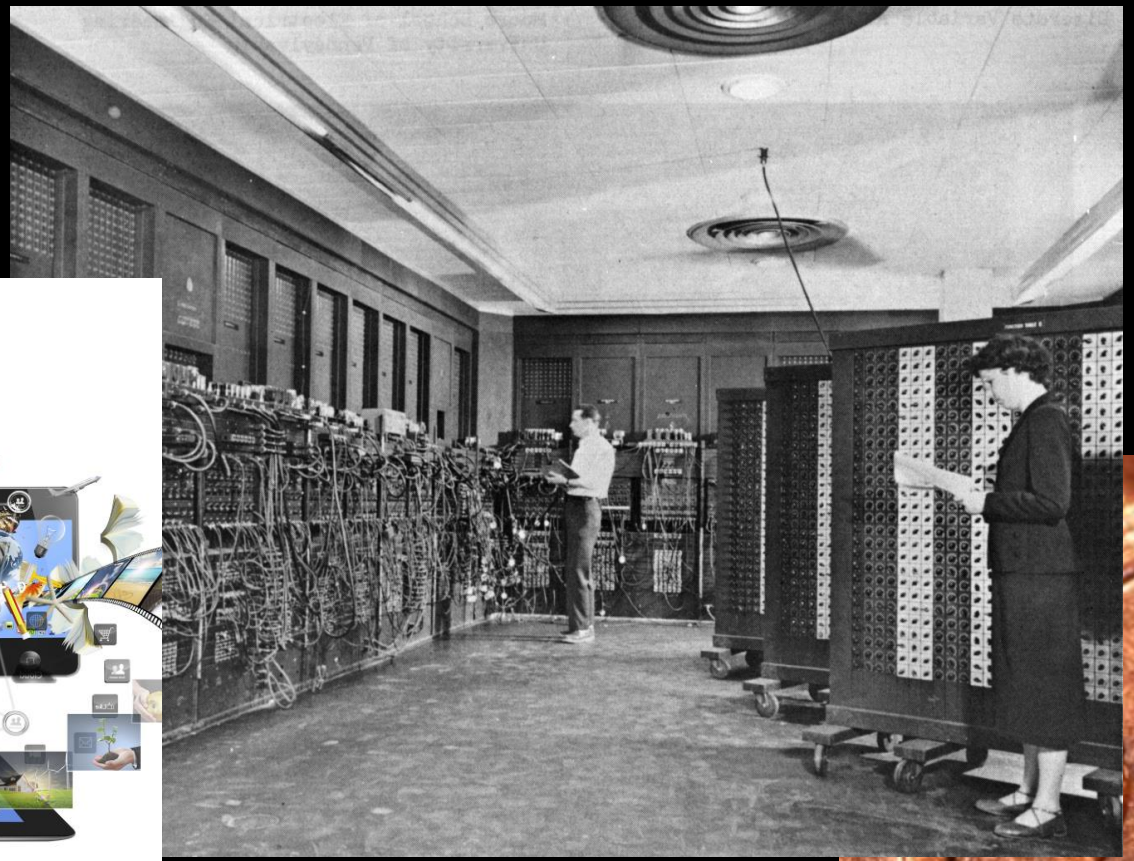
$$= 1$$

$$\theta_{i,j} = \frac{t=1}{\sum_{t=1}^{T} P(x_{t-1} = j, Y, \theta^{old})}$$

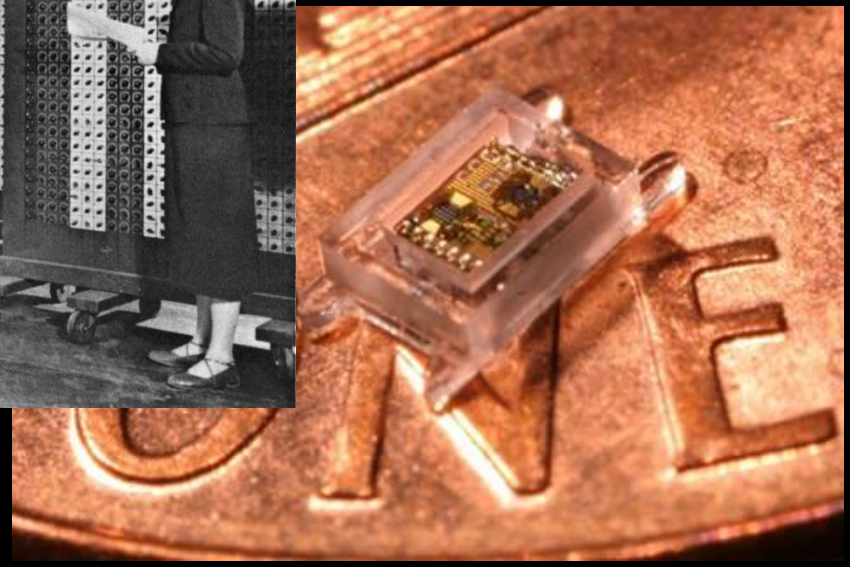
We will discuss this part ->

http://www.memegasms.com/media/created/hyo3gj.jpg

WHY IS ASR READY NOW?



https://en.wikipedia.org/wiki/ENIAC#/media/File:Eniac.jpg



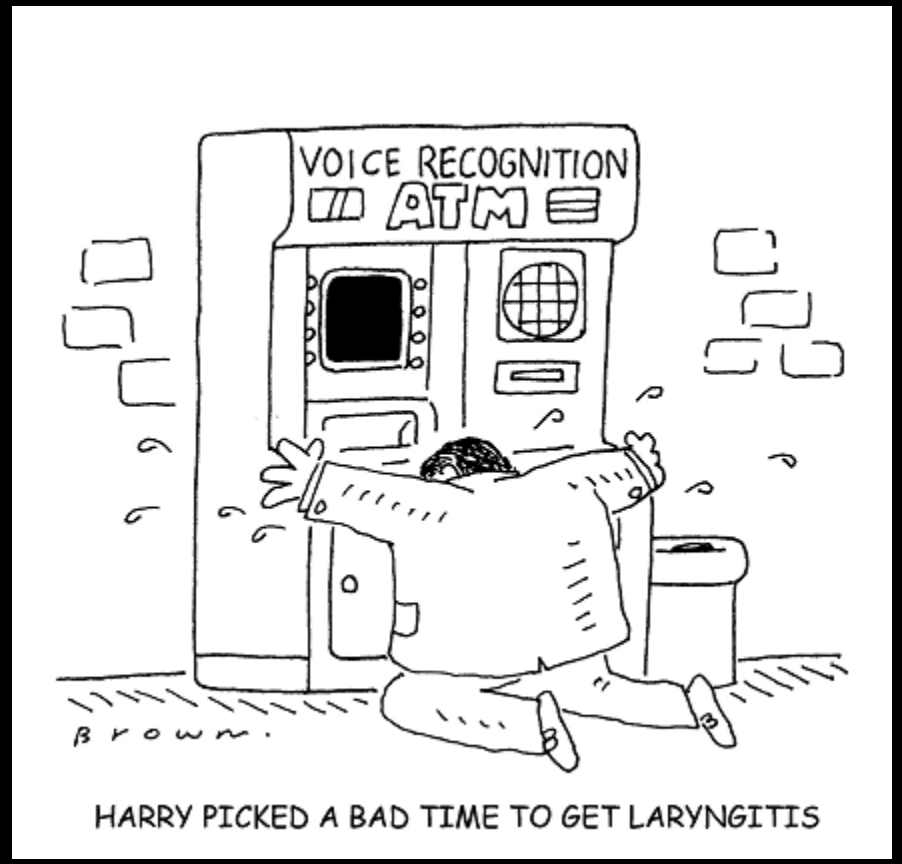
http://www.gelsolution.com.br/site/wp-content/uploads/2014/11/site_cloud-computing-1.jpg

http://media4.s-nbcnews.com/j/MSNBC/Components/Photo/_new/110222_tech_tiny-computer.grid-6x2.jpg

SOME PRIOR WORK ON ASR TOOLS FOR DHH

- **Non-Real-Time:**
  - Captioning online lecture videos [Shiver and Wolfe]
- **Semi-Automated Real-Time:**
  - Captioning classroom lectures with human overseers [Gaur *et al.*]
  - Crowd caption correction during meetings [Harrington and Vanderheiden]
- **Fully Automated Real-Time:**
  - Augmented Reality glasses [Mirzaei *et al.*]

ASR FAIL



http://my-ecoach.com/online/resources/13023/voice-recognition-cartoon.png

http://media.techeblog.com/images/youtube_fail.jpg

COMPREHENSION IS IMPORTANT

- "knowing the **context** and searching for **keywords** are **essential steps** to build their capacity of **understanding**"

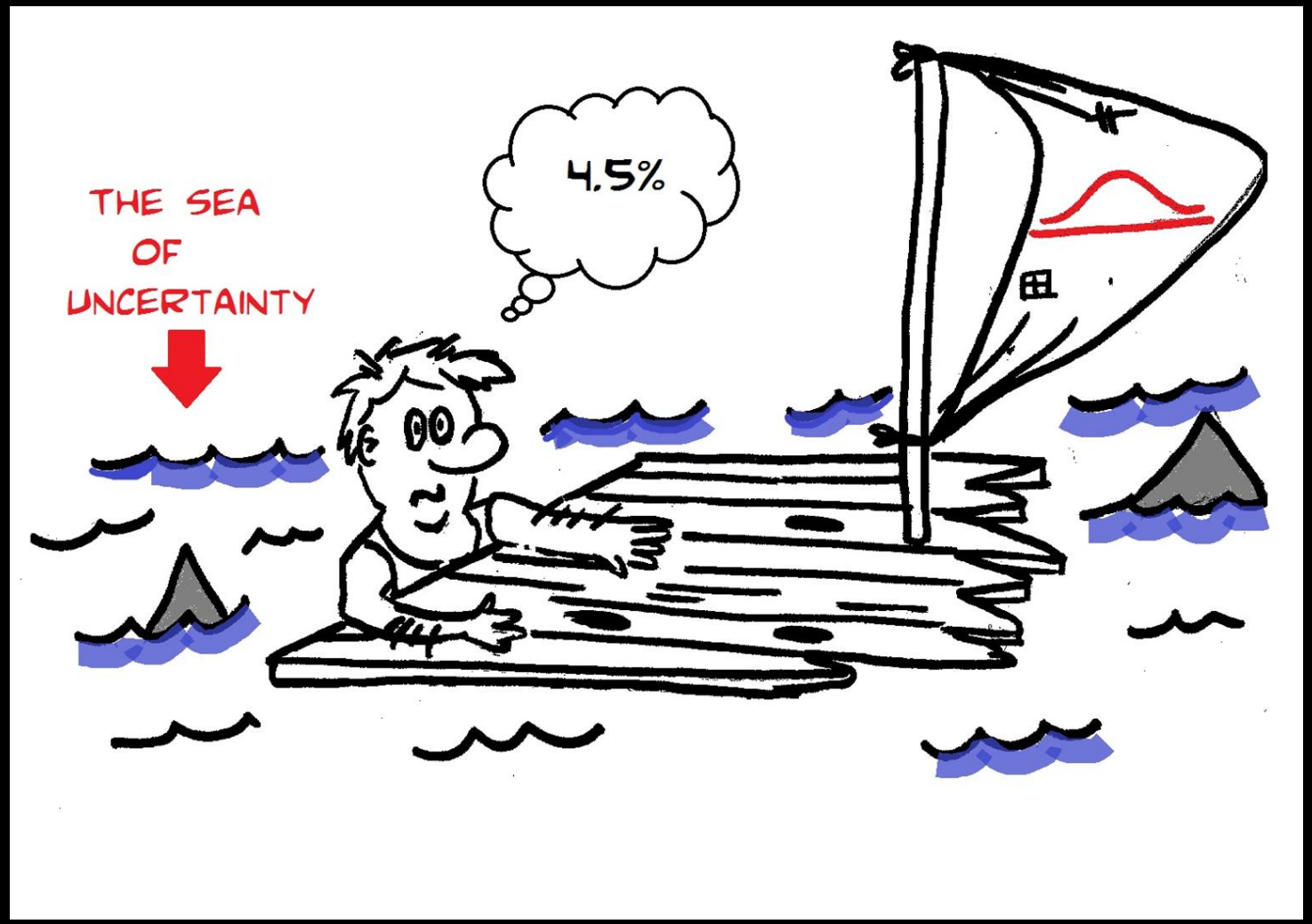
~participant in *Qualitative investigation of the display of speech recognition results for communication with deaf people* [Agnès Piquard-Kipffer *et al.*]

Thus DHH users might be able to benefit from imperfect ASR technologies as long as they see comprehensible and correct keywords in the output!

OUR RESEARCH FOCUS

- Live one-on-one meeting between a DHH individual and a hearing person using ASR
- Different approach than tools designed for classrooms or corporate meeting rooms
- Utilize ASR in a reduced-noise environment and provide the speaker with feedback so that they can change how they speak for improved results

WHAT IS CONFIDENCE?



http://www.patientdriven.org/wp-content/uploads/2014/01/Sea-of-uncertainty.jpg

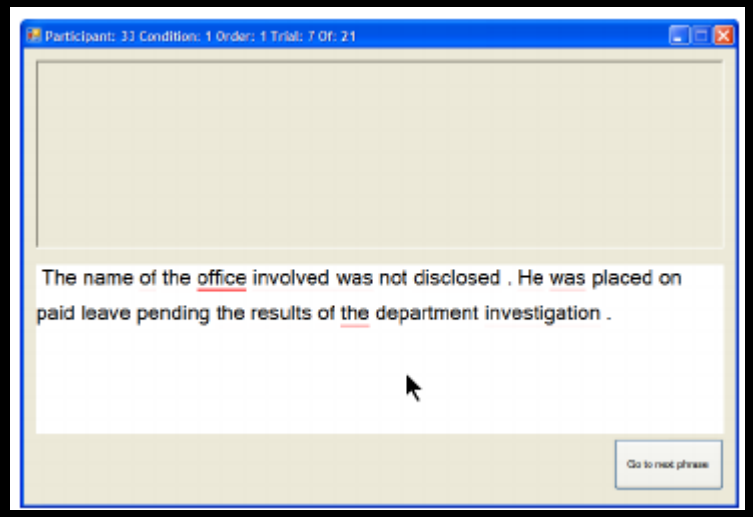
PRIOR WORK ON DISPLAYING CONFIDENCE

- **Font Change** – Piquard-Kipffer *et al.*
- **Font Color** – Shiver and Wolfe
- **Underlining** – Vertanen and Kristensson

	words/syllables tagged as ***correct*** are displayed in **bold**
words tagged as *incorrect* are displayed into **orthographic** mode	**je voudrais être** li vré qu'on bien **ça** kou te
words tagged as *incorrect* are displayed into **pseudo-phonetic** mode	**je voudrais être** l i v r é k on b y in **ça** k ou t e

Piquard-Kipffer *et al.*

Shiver and Wolfe

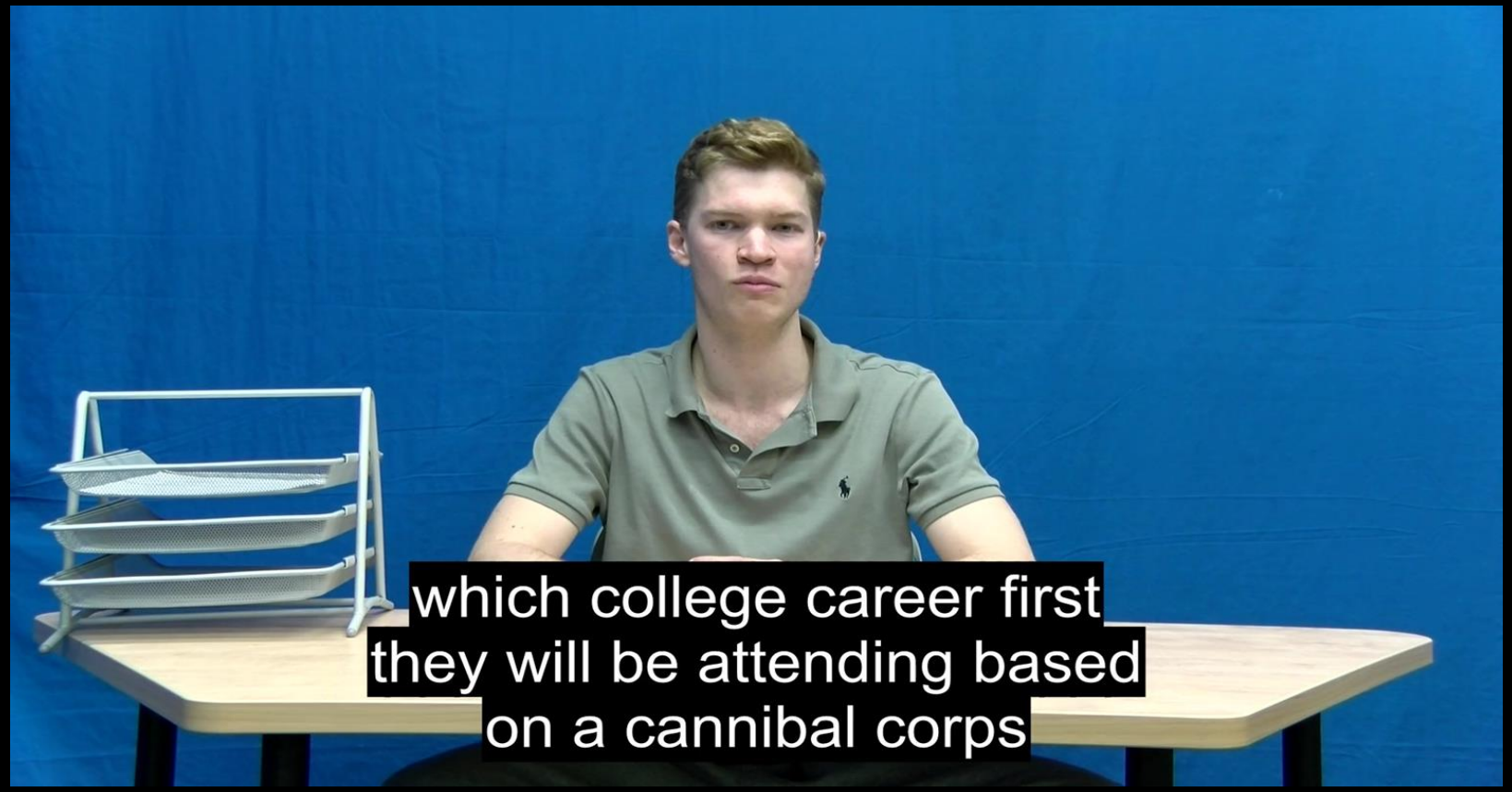


Vertanen and Kristensson

ALTERNATIVE CAPTIONING METHODS

- Font Change
- Font Color
- Underlining
- Colored Borders
- Dynamic Positioning
- Dynamic Size
- Emoji
- Removal of Text
- Syllables
- Text Spacing
- Tracked Display
- Transparency

OUR PROTOTYPE TOOL



Actual Script: “which college career fairs
they will be attending based
on the candidate requirements”

12 CAPTION MARKUP STYLES

No Change (no_change)

Bold on Confident (bold_c)

Bold on Uncertain (bold_u)

Green on Confident (color_c)

Red on Uncertain (color_u)

Delete on Uncertain (del_u)

Italics on Uncertain (it_u)

Range of Gray Color (r_gray)

Range of Font Size (r_size)

Smaller Size on Uncertain (size_u)

Underline on Uncertain (ul_u)

Underline and Gray on Uncertain (ul_gray_u)

PILOT METHODOLOGY

- Gather DHH participants from RIT/NTID and have them participate via an HTML-based experiment-presentation software
- Pre-experiment: demographics/general questions
- View short videos simulating a one-on-one business meeting with a colleague (with ASR captioning)
  - Divide the meeting into 12 “paragraphs” and apply different markup styles on the captioning for a “latin squares” within-subjects experiment
  - Measure participants' preference of the caption display style and validate their comprehension of the information content
- Post-experiment: ask the participant to rank the best caption appearance methods

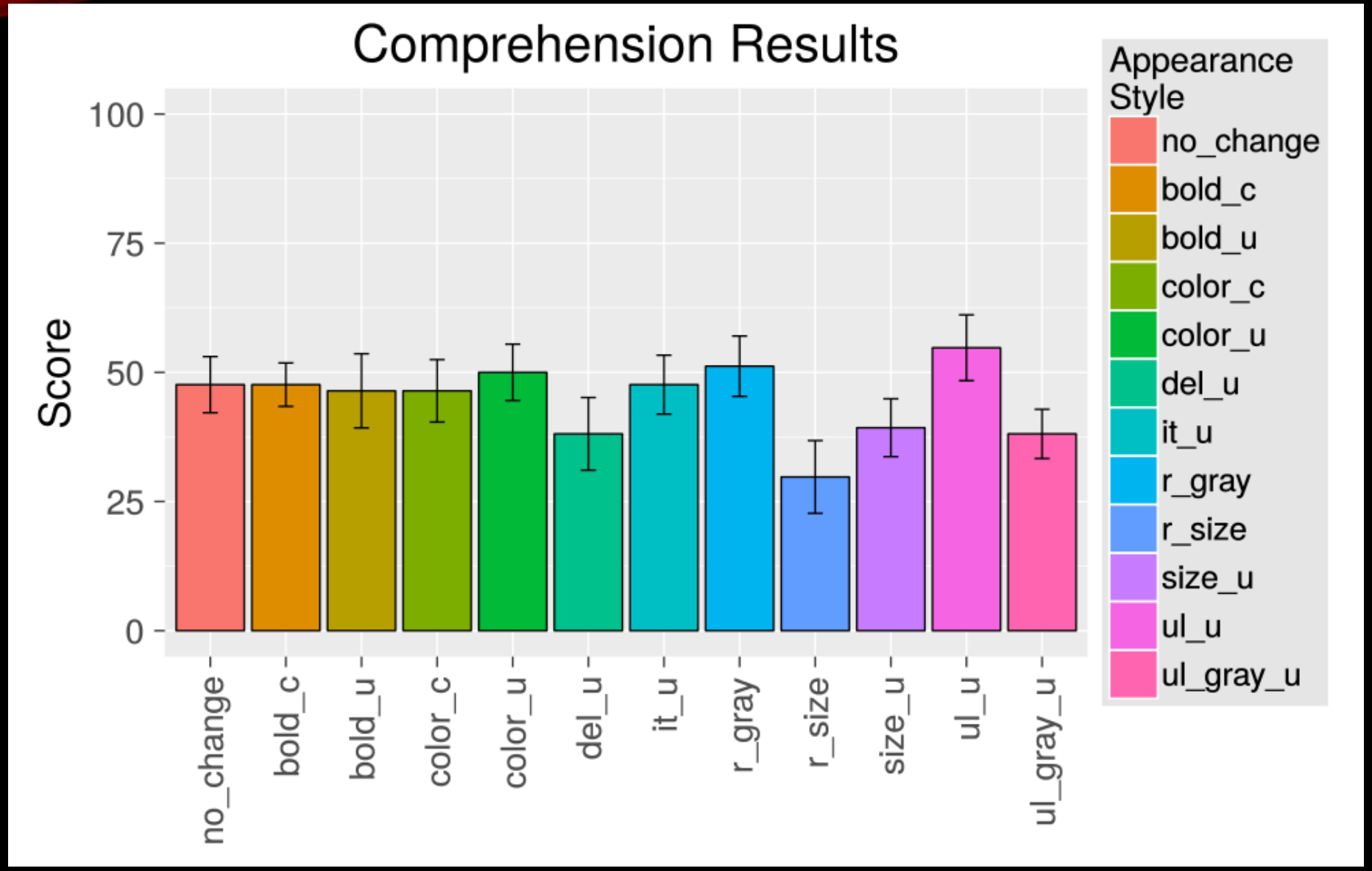
STIMULI PREPARATION

PILOT TEST

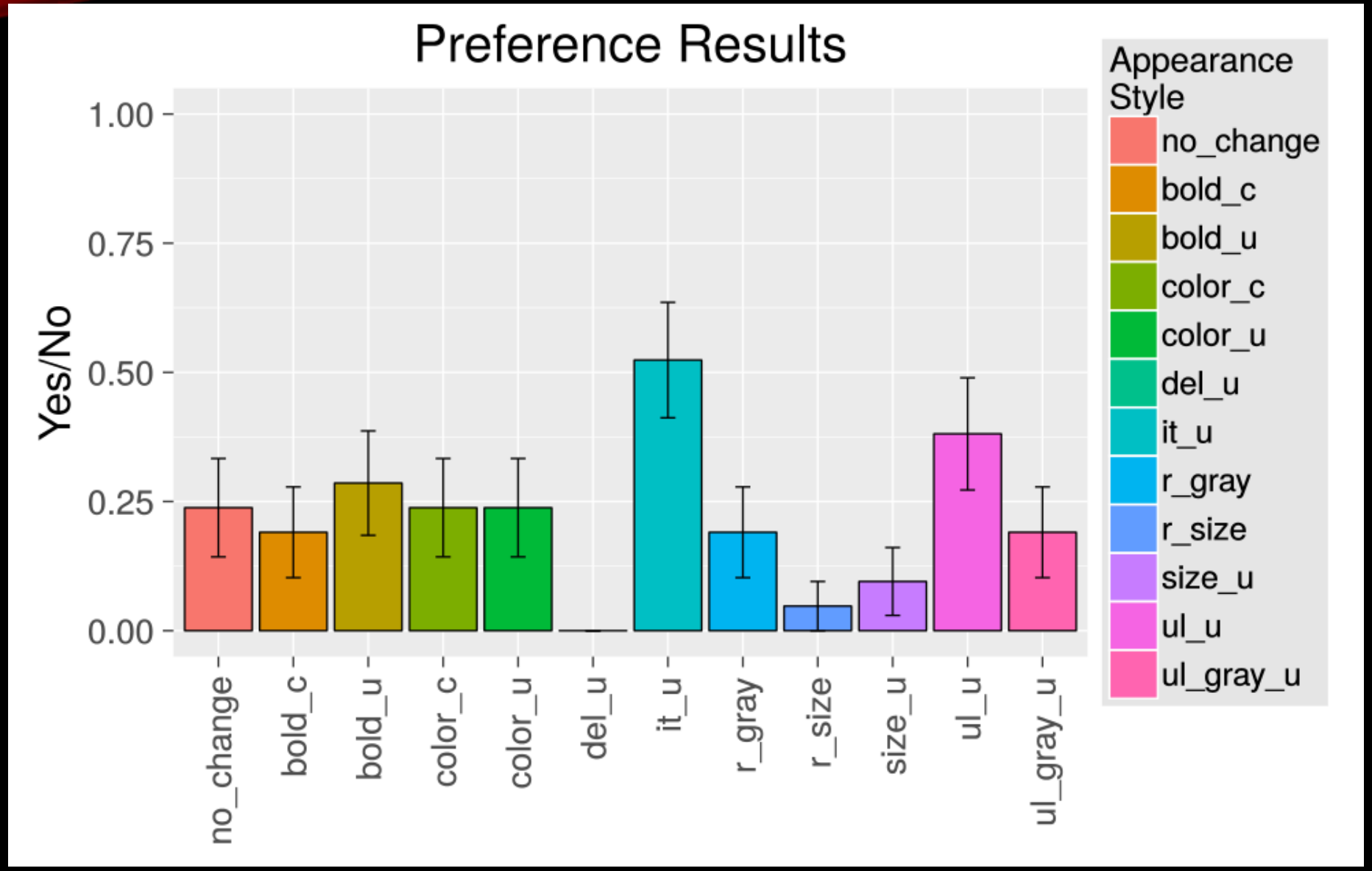
- We executed a preliminary study during 2016
- 21 DHH students from NTID/RIT and 11 hearing students from RIT
- Initial results matched our expectations which served as a 'sanity check' on the design
- We didn’t obtain significant results but it was invaluable in helping us design the follow-up experiment to occur this spring semester

https://upload.wikimedia.org/wikipedia/commons/7/7a/Chuck_Yeager.jpg

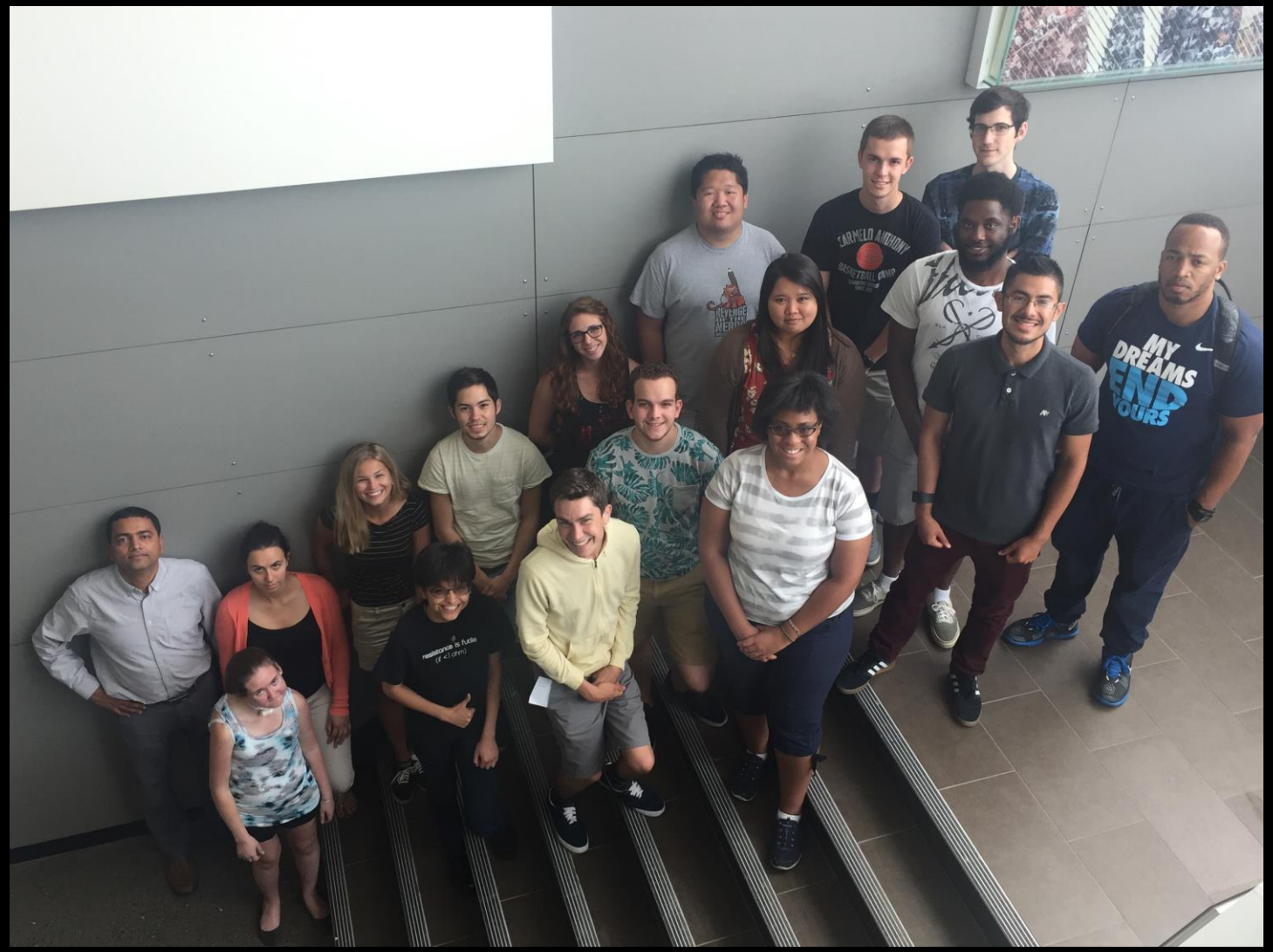
DHH PILOT RESULTS



DHH PILOT RESULTS



AMI REU



- Accessible Multimodal Interfaces
- Research Experiences for Undergraduates - Summer 2016
- Dr. Raja Kushalnagar led the program
- Dr. Mike Stinson and Dr. Matt Huenerfauth mentored 4 students:
  - Paul Bayruns
  - Kevin Rathbun
  - Daniel Saavedra
  - Abigail Spring

DISCUSSION OF THE PILOT STUDIES

- Many researchers reported that it is possible to utilize the confidence values from the ASR engine
- DHH users would need to adapt to an imperfect ASR world for a while
- Standardizing the display of confidence could help to reduce the mental load of the DHH user

FUTURE WORK

- We are in the planning stages of a follow-up study for this spring semester with a larger number of DHH participants to further explore the markup's influence on comprehension
- Dr. Stinson's team is working on an Android-based software that implements our initial findings so we can test it with participants in a more realistic environment instead of a mock meeting
- Sushant is currently investigating how we can modify the ASR's choice of words to improve comprehension by DHH users
- Christopher is planning a study to investigate how captioning styles influence the hearing speaker's behavior
- We plan to continue working with Dr. Kushalnagar during this summer's AMI session to follow up on the eye-tracking results and elicit additional research ideas

http://www.clipartkid.com/images/497/under-construction-signs-clipart-best-OmYeS3-clipart.jpeg

http://thechronicleherald.ca/sites/default/files/imagecache/ch_article_main_image/articles/questions.jpg