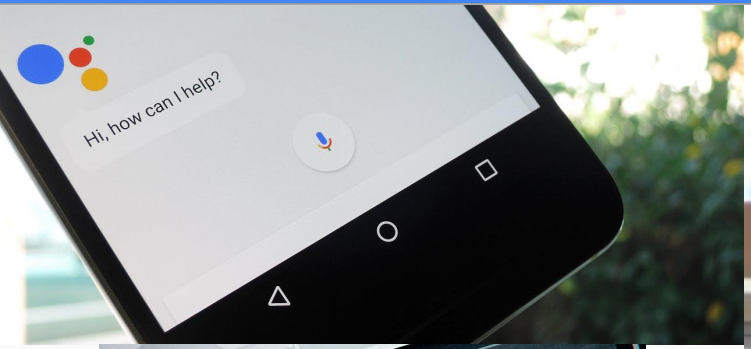
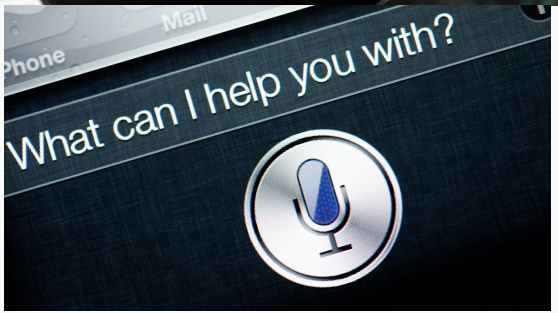
# Automatic Speech Recognition: Relationship between Text Readability and Word Error Rate

Abraham Glasser
Rochester Institute of Technology
Rochester, New York


32nd CSUN Assistive Technology Conference
San Diego, California

# ASR: What is it?

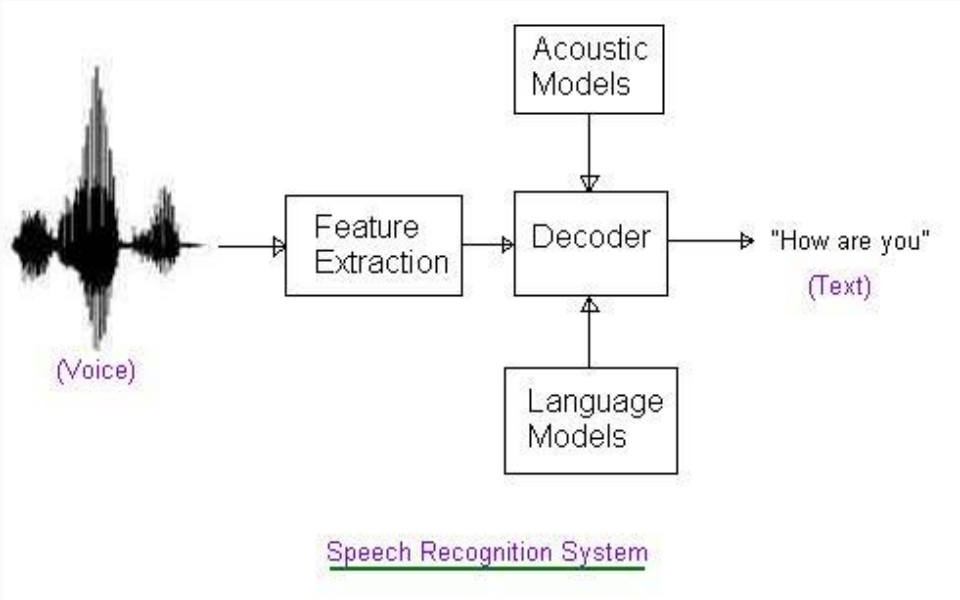




https://goo.gl/A3Za2A, https://goo.gl/SrwCcg, https://goo.gl/KZ0vpu

# ASR: How does it work?


https://goo.gl/A5rsdt, https://goo.gl/LwiqzJ

# Word Error Rate vs Flesch-Kincaid Readability

- Goal of this project is to analyze the relationship between ASR word error rates and the complexity of the vocabulary in texts
- Also this project talks about why such a relationship would exist and the reasoning behind an ASR error
- This project is useful because
  - We can use it to see how an ASR would perform in different environments
  - Eliminates/Confirms the factor of complexity when understanding why an ASR makes errors

# Background Work: Readability Measure

Flesch-Kincaid Readability Measure

- Average length of words

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

- Average length of sentences

| Score | School Level |
|---|---|
| 90.0–100.0 | 5th grade |
| 80.0–90.0 | 6th grade |
| 70.0–80.0 | 7th grade |
| 60.0–70.0 | 8th & 9th grade |
| 50.0–60.0 | 10th to 12th grade |
| 30.0–50.0 | college |
| 0.0–30.0 | college graduate |

# Background Work: ASR Accuracy Improvement

- Acoustics
  - Improving Audio Quality
  - Reducing Speech Reverberation

- Language
  - Making better algorithms in choosing words
  - Trying different approaches in language models

# Methodology

- Scripts
  - Collected and adapted from various online sources about American History
  - Average word count of ~230
  - F-K Scores ranged from 9.5 to 70.4
  - Read aloud in natural human speech
- ASR
  - IBM Watson Speech to Text API through IBM Bluemix
  - Microsoft Bing Speech API through Microsoft Azure

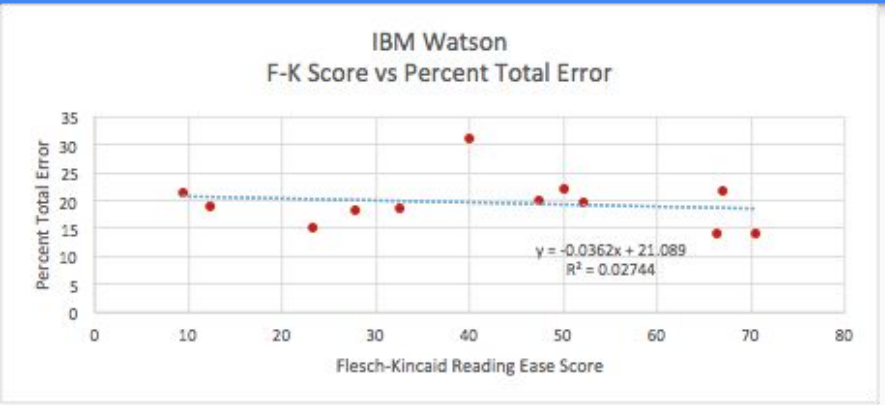
https://www.ibm.com/watson/developercloud/speech-to-text.html
https://www.microsoft.com/cognitive-services/en-us/speech-api

# Technology cont.

- Word Error Rate
  - NIST (National Institute of Standards and Technology)
    - Speech Recognition Scoring Toolkit (SCTK) Version 2.4.0
  - Looks at Substitutions, Deletions, Insertions, etc
  - Calculates “Percent Total Error”
    - Percent Total Error = 100 - Word Accuracy = Word Error Rate

http://www.itl.nist.gov/iad/mig//tools/

# Watson ASR vs F-K Score



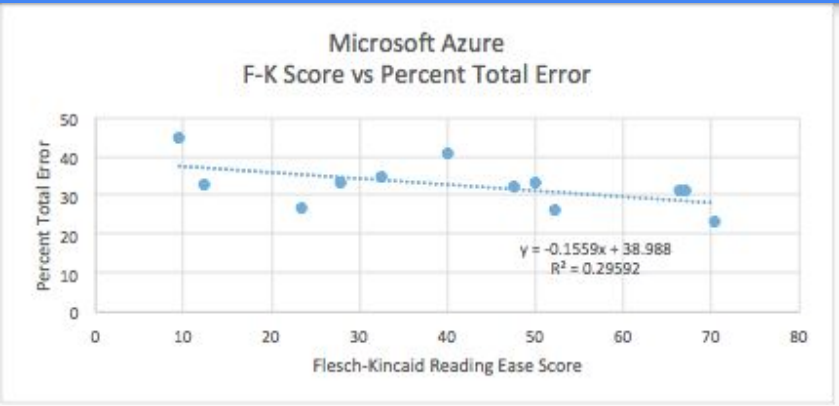

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.165663221 |
| R Square | 0.027444303 |
| Adjusted R Square | -0.069811267 |
| Standard Error | 4.72932466 |
| Observations | 12 |

The R^2 is ~.027, which means that the relationship is very weak

The p-value is much greater than .05, which means that it is not a statistically significant relationship

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 21.08941917 | 3.146765011 | 6.701936464 | 5.35168E-05 | 14.07798979 | 28.10084855 | 14.07798979 | 28.10084855 |
| Flesch-Kincaid Reading Ease Score | -0.036235925 | 0.068213525 | -0.531213203 | 0.606867719 | -0.188225131 | 0.115753281 | -0.188225131 | 0.115753281 |

# Azure ASR vs F-K Score



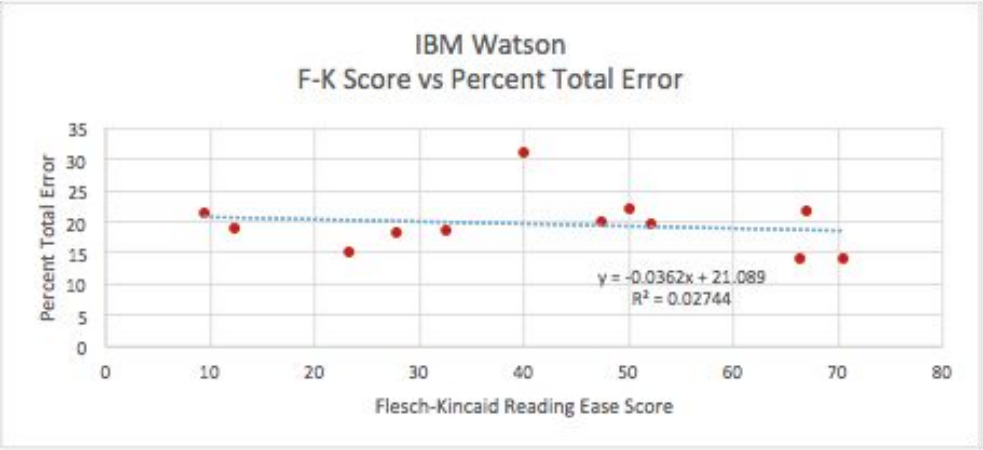

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.543987968 |
| R Square | 0.29592291 |
| Adjusted R Square | 0.225515201 |
| Standard Error | 5.27207669 |
| Observations | 12 |

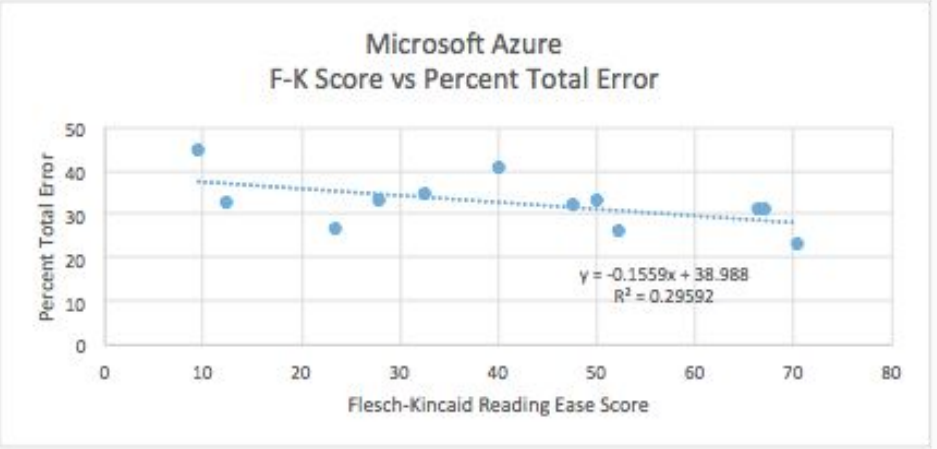
The R^2 is ~.3, which means that the relationship is weak, there is a weak negative relationship

The p-value is a bit more than .05, which means that it is not a statistically significant relationship

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 38.98785046 | 3.507897566 | 11.114307 | 5.98764E-07 | 31.17176761 | 46.80393332 | 31.17176761 | 46.80393332 |
| Flesch-Kincaid Reading Ease Score | -0.155895031 | 0.076041922 | -2.050119553 | 0.067494792 | -0.325326991 | 0.01353693 | -0.325326991 | 0.01353693 |

# Comparison


IBM Watson
F-K Score vs Percent Total Error
Percent Total Error
Flesch-Kincaid Reading Ease Score
y = -0.0362x + 21.089
R² = 0.02744
0
5
10
15
20
25
30
35
0
10
20
30
40
50
60
70
80


Microsoft Azure
F-K Score vs Percent Total Error
Percent Total Error
Flesch-Kincaid Reading Ease Score
y = -0.1559x + 38.988
R² = 0.29592
0
10
20
30
40
50
0
10
20
30
40
50
60
70
80

# Conclusions

| ASR System | IBM Watson | Microsoft Azure |
|---|---|---|
| Statistical Significance | Not statistically significant<br>(.607 >> .05) | Almost statistically significant<br>(.067 > .05) |
| Relationship strength | Very weak | Weak |

- It seems that lower F-K scores (more complex vocabulary) does have an impact on the error that an ASR will have. However, correlation analysis shows that there is a weak relationship.
- The graphs both show a small trend showing a decrease in WER with an increase in F-K scores.

# Conclusions cont.

- Interesting small difference between the two ASR's
- While one ASR was more accurate, the other one was more "readable" (had punctuation, numerals, etc.)
- ASR has been focused more on short voice commands or sentences

| Microsoft ASR | Original | Watson ASR |
| --- | --- | --- |
| Jamericans, a whole war was a civil war in the United States fought from **1861 to 1865.** The union face sessionista and **11**. Southern states known as the Confederate States of America. | The American Civil War was a civil war in the United States fought from 1861 to 1865. The Union faced secessionists in **eleven** Southern states known as the Confederate States of America. | The American Civil War was a civil war in the United States fought from **eighteen sixty one to eighteen sixty five** the union effaced secessionists in eleven southern states known as the Confederate states of America. |

# Future Work

- This was a preliminary study

- Try different fields
  - Biology, psychology, physics, etc.
- Repeat data with a different human to read aloud the texts
- Try with more ASR systems

# Summary and Future Work

1. In general, for American history content, less complex words → less error
   a. Has yet to be tested with subjects such as biology, psychology, mathematics, engineering
2. It is important for ASR to be accurate and easy on the eyes to read
   a. Punctuation, line breaks, numerals
3. Factors impacting using ASR in the real world (classrooms, meetings)
   a. Voice/speed, background noise, disfluencies in speech
   b. These were not included in this research