The Crux of Voice (In)Security: A Brain Study of Speaker Legitimacy Detection

Ajaya Neupane[1,2], Nitesh Saxena[2], Leanne Hirshfield[3], Sarah Elaine Bratt[3]

[1]University of California Riverside

[2]University of Alabama at Birmingham

[3]Syracuse University

Motivation

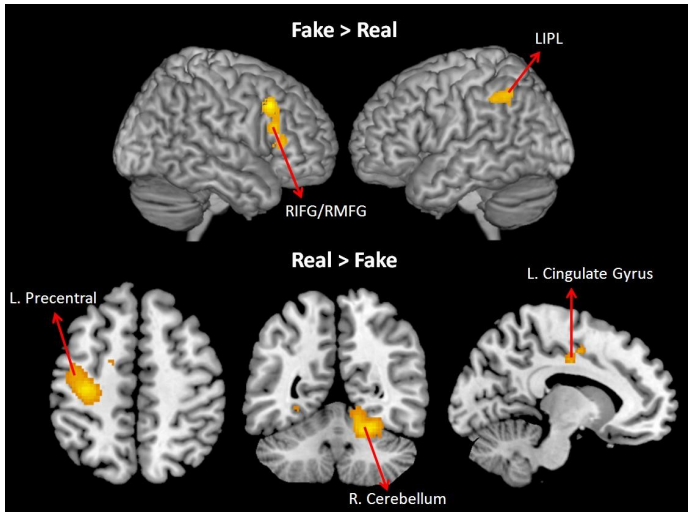
- We leave voice traces behind
- How difficult is it to make a machine talk like you?
  - Off-the-shelf speech morphing tools (e.g., Lyrebird, Festvox) can be used to generate the spoofed voices of people
- Voice is used as a biometrics
  - Voice-based user authentication systems
- Voice scams are run to fool users
  - Examples, grandparent scam, post malicious message in social media

Premise

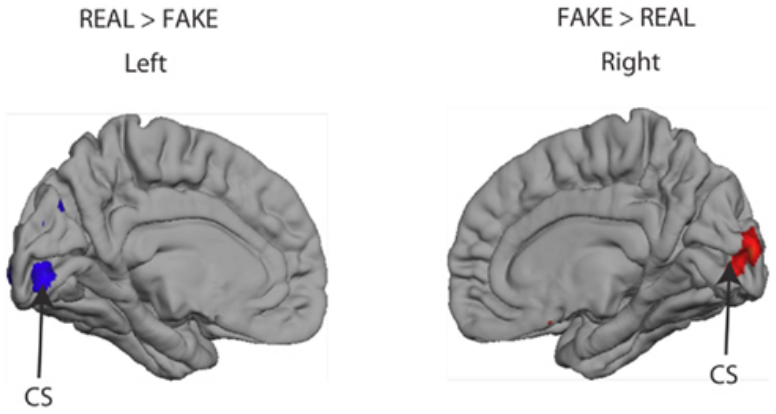
- Human-based speaker verification systems are vulnerable to voice morphing [Shirvanian et al.; CCS 2014, Mukhopadhya et al.; ESORICS 2015]
- In this study, we focus on analyzing why human users are vulnerable
- Studies have reported differences in neural activities when users are subject to real-fake artifact detection
  - Website legitimacy detection [Neupane et al.;NDSS'14, CCS'15, WWW'17]
  - Image legitimacy detection [Huang et al.; Frontiers of Human Neuroscience'11]

Hypothesis

The activation in frontopolar and temporoparietal areas will be high when participants are listening to the morphed voice of a speaker compared to the original voice of the speaker



[Neupane et al.; NDSS 2014]



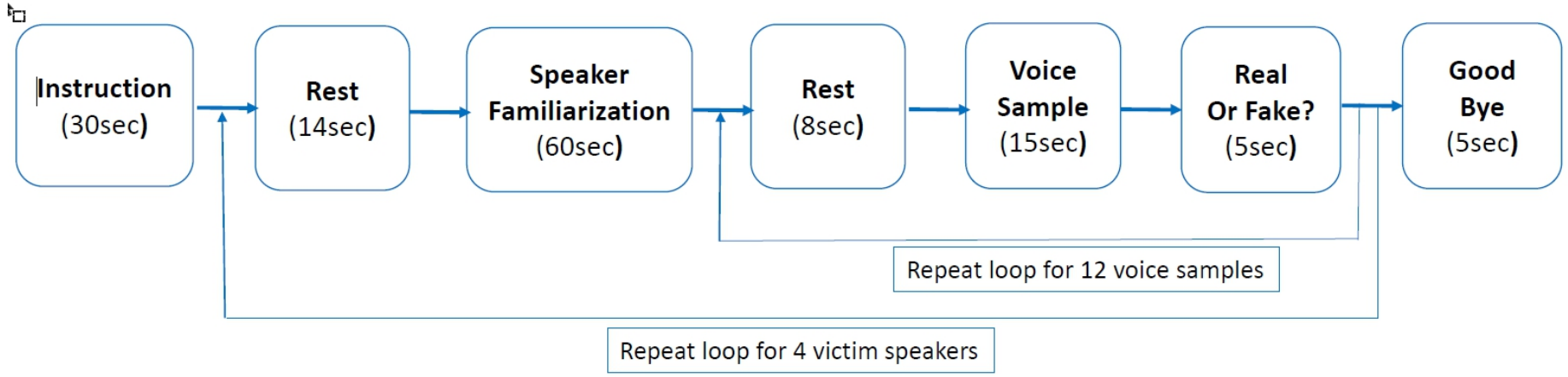
Huang et al.; Frontiers in Human Neuroscience'2011

Task Overview

- Speaker Legitimacy Detection
  - The act of identifying whether the given voice sample is real or fake
- Voice samples (Victim speakers)
  - Familiar voices - Morgan Freeman, Oprah Winfrey
  - Briefly familiar voices - Two mechanical turk users
- Fake of voices used in our study
  - Different speaker voice generated using mechanical turk users
  - Morphed speaker voice generated using CMU Festvox convertor

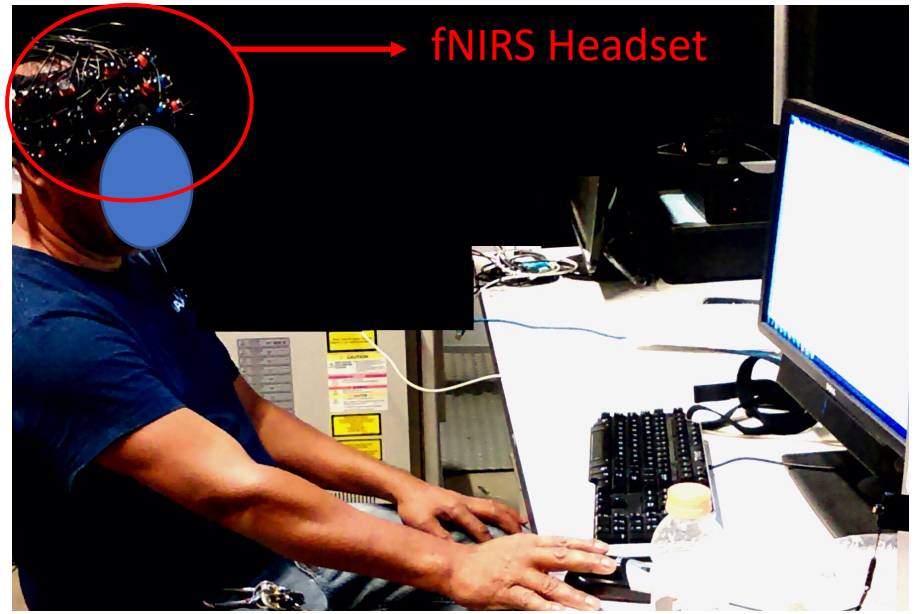
Task Design

- Familiarize participants with the original speakers' voice for 60 second
- Play a sample each of real or a different or a morphed speakers' voice
  - For each speaker we played 4 real, 4 morphed and 4 different voices
- We ask participants to identify real and fake voices



Study Set-up

- Use fNIRS (functional Near-Infrared Spectroscopy)
  - BOLD principle
  - Measures oxy-Hb and deoxy-Hb
  - Better spatial resolution than EEG
  - Better temporal resolution than fMRI
- Laptop to display stimuli
  - Recorded behavioral data
- Desktop to record brain activities
  - Recorded fNIRS measurements



Data Collection

- Study IRB approved
- Recruited twenty healthy participants from the broader university community (including students and staff)
- All English speaking participants
- 10 were male and 10 were female
- Age range of 19-36 years with a mean age of 24.5 years

Data Processing

- Removed high frequency noises and artifacts
- Averaged neural activities measured when participants were listening to a voice sample
- Compared the neural activities when participants were listening to real, morphed, and different speakers' voices

Data Analysis

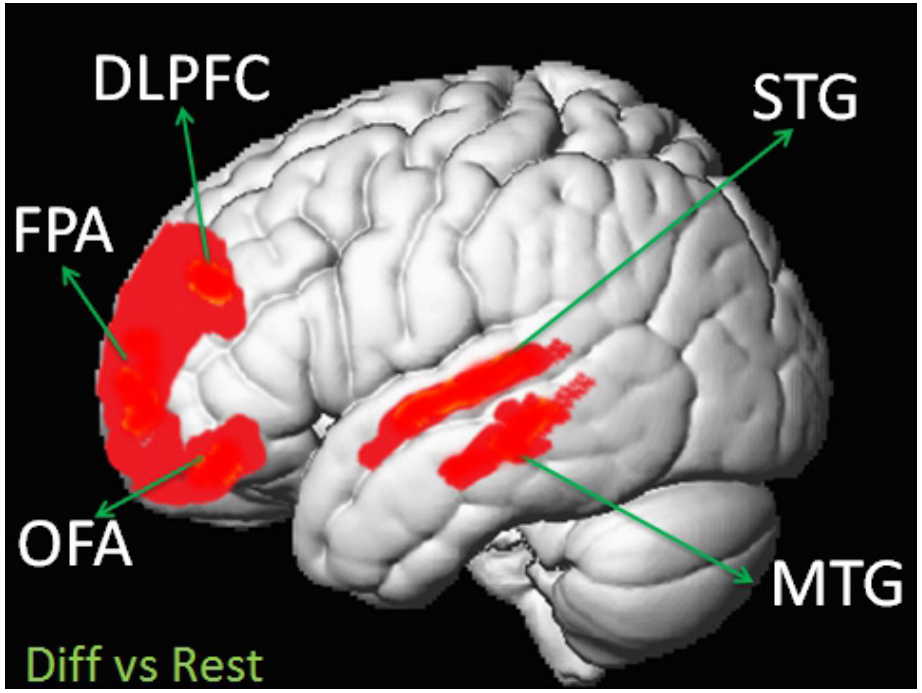
- We evaluated activities at five specific areas based on findings of previous studies
  - Dorsolateral prefrontal cortex, orbitofrontal gyrus, frontopolar area
  - Superior temporal gyrus and middle temporal gyrus
- Statistical Analysis
  - Non-normal distribution
  - Friedman's Test – comparison across all groups
  - Wilcoxon Signed Rank Test – comparison across specific groups
  - Bonferroni Correction - correction for multiple tests

Functions of Brain Areas

ROI Name	Acronym	Brodmann Area #	Functionality
Dorsolateral Prefrontal Cortex	DLPFC	9	Working memory, attention
FrontoPolar Area	FPA	10	Memory recall, executive functions
Superior Temporal Gyrus	STG	22	Primary auditory cortex, auditory processing
Middle Temporal Gyrus	MTG	21	Recognition of known faces
Orbitofrontal Area	OFA	11	Cognitive processing, decision making, trust

Results: Are participants trying?

- Original vs. Rest
- Morphed vs. Rest
- Different vs. Rest



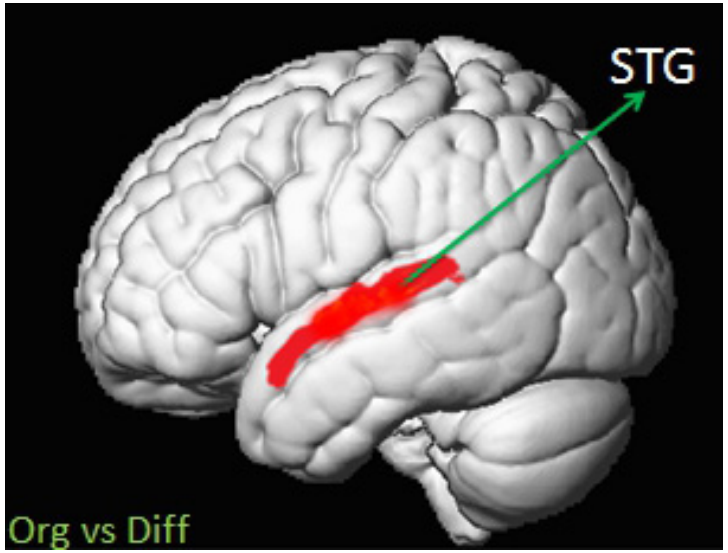
Results: Original vs. Morphed Voice

- **Observation:** We did not observe any statistically significant differences in any of the five areas we considered for analysis, irrespective of the gender type

- **Interpretation**: Morphed voices may have sounded identical to the original voices

Results: Original vs. Different Speaker's voice

- At superior temporal gyrus, we observed that the neural activation for original speakers voice was higher than different speaker's voice
  - Results similar to study by Bethmann et al. [PloS one'12]
- We did not observe statistically significant differences in other areas of brain
- Similar observation for both genders



Task Performance

Trial	Acc	Prec	Rec	FM	RTime
	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)
Original	82.1 (16.6)	50.63 (12.5)	83.2 (16.1)	61.31 (9.0)	2.57 (0.5)
Morph	42.8 (24.1)	46.71 (19.7)	42.8 (24.1)	43.40 (19.8)	2.58 (0.5)
Different	67.2 (21.5)	47.82 (16.0)	68.1 (21.2)	55.82 (16.0)	2.51 (0.5)
Average	64.2 (11.5)	48.39 (15.3)	64.7 (20.7)	53.51 (15.0)	2.54 (0.5)

- Observations
  - Overall accuracy of correctly identifying a voice 64%
  - Participants reported 58% of morphed speakers as real speakers
  - Participants reported 33% of different speakers as real speakers

Machine Learning on Brain Data

- How much can we learn from neural activities of human users' on the legitimacy of the voices they listen to?
- Normalized fNIRS (oxy-Hb and deoxy-Hb) data
- Extracted Features - maximum, minimum, average, standard deviation, slope, variation, skew, and kurtosis
- Off-the-shelf machine learning algorithms including J48, Random Forest, Neural Networks, Support Vector Machines, Naïve Bayes

ML: Original vs Morphed Voice

Classifier	Precision Mean (Std)	Recall Mean (Std)	F-Measure Mean (std)
Random Tree	49.5 (9.5)	48.8 (10.9)	48.9 (9.7)
Logistic	49.4 (11.2)	50.0 (11.9)	49.4 (10.9)
J48	48.8 (10.7)	48.8 (12.1)	48.6 (11.2)
NaiveBayes	46.7 (10.1)	43.8 (11.2)	44.8 (9.5)
Multilayer Perceptron	50.0 (11.3)	48.8 (11.1)	49.0 (10.3)
LMT	48.3 (11.1)	54.4 (12.4)	50.8 (10.9)
Simple Logistic	49.4 (10.3)	59.1 (13.5)	53.2 (10.1)
SMO	49.0 (12.6)	47.2 (12.9)	47.8 (12.0)
Random Forest	47.1 (9.5)	52.8 (11.4)	49.6 (9.8)

Observation:

- Best F-measure of identifying the voice of morphed speaker vs. original speaker was 53%

Discussion

- Human may be inherently incapable to distinguish morphed voices
- Training human may improve their performance
  - Make users aware of voice morphing attacks
- Developing technical solutions to assist the users
- The voice morphing technology may be ready to serve those who have lost their voices

Study Limitations

- Lab environment
- Repeated trials
- Young participants
- Resolution of fNIRS

Conclusions

- Explored voice security through the lens of neuroscience and neuroimaging
- No significant differences when users were subject to real vs. morphed voices
- Low behavioral performance in identifying morphed voices
- Users' may be biologically susceptible to morphing attacks, and only proper training may help

Thank You

Q&A

Demographics

Participant Size N = 20	
Gender(%)	
Male	50%
Female	50%
Age(%)	
18-22	20%
22-26	55%
27-31	20%
31+	5%
Handedness(%)	
Right-Handed	90%
Left-Handed	10%
Background(%)	
High School	10%
Bachelor's	20%
Masters	55%
Doctorate	10%
Others	15%

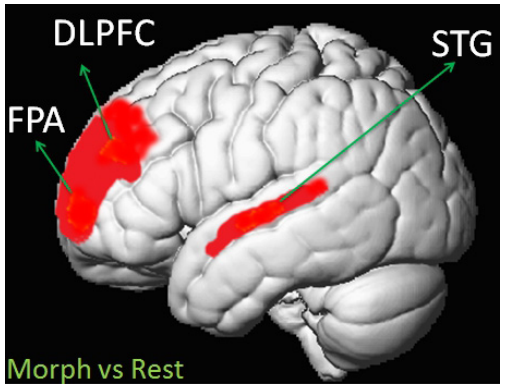
Premise

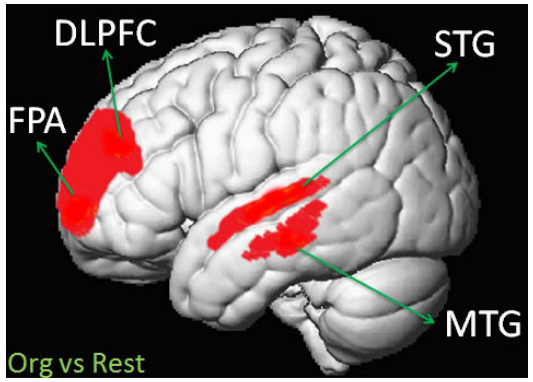
- What are the consequences?
- Voice is used as a biometrics
  - e.g., voice-based user authentication systems
- Voice makes us known to people
  - e.g., attacking arbitrary speech contexts

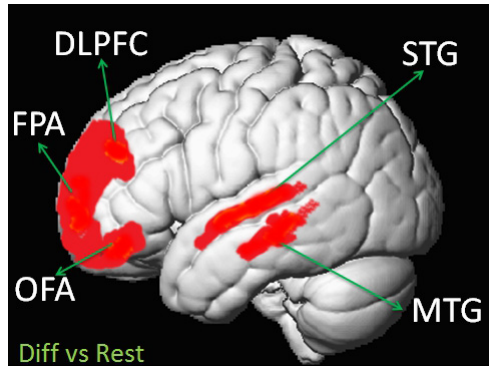
Background

- Previous studies reported differences in neural activities between real and fake artifacts
  - Observed neural differences between real and fake paintings
- Our Studies [Study I, NDSS'14; Study II, CCS'15; Study III, WWW'17]
- Asked participants to identify real and fake websites
- Differences in neural activities when participants viewed real and fake websites
- We want to see if these differences exists when people are listening to real and fake voices of a speaker

Neural Results: Rest vs. Voice Trials

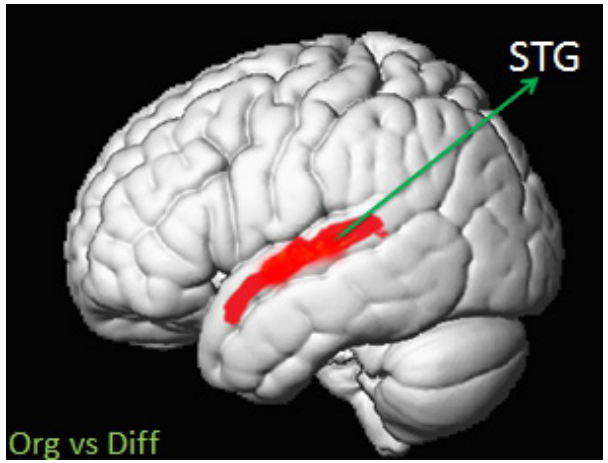






- **Observation:** There were more neural activity at the dorsolateral prefrontal cortex (DLPFC), frontopolar area (FPA), orbitofrontal area (OFA), superior temporal gyrus (STG), and middle temporal gyrus (MTG) for voice trails compared to the rest trials

Neural Results: Original vs. Fake Voices



- **Original vs Different Voices**
  - **Observation:** At superior temporal gyrus neural activities for original speakers voice> different speaker's voice
  - **Interpretation**: It shows known speakers voice generates higher activation in superior temporal gyrus compared to the unknown voices

- **Original vs Morphed Voices**
  - **Observation:** No statistically significant differences
  - **Interpretation**: Morphed voices may have sounded identical to the original voices