Voice Liveness Detection for Voice Assistants using Ear Canal Pressure

Jiacheng Shang
Dept. of Computer Science
Montclair State University

Jie Wu
Center for Networked Computing
Dept. of Computer and Info. Sciences
Temple University

MONTCLAIR STATE UNIVERSITY

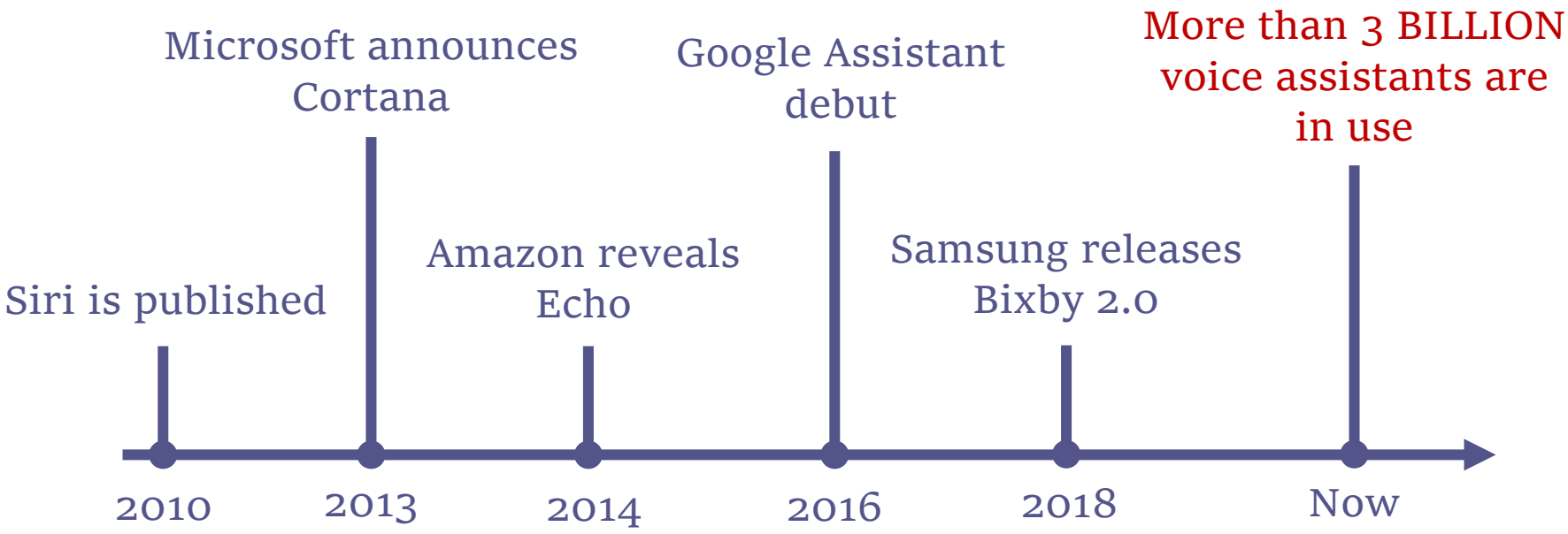
Voice Assistants on Smart Speakers

- A digital assistant that uses voice recognition, speech synthesis, and natural language processing (NLP) to provide a service.

MONTCLAIR STATE UNIVERSITY

Voice Assistants on Smart Speakers

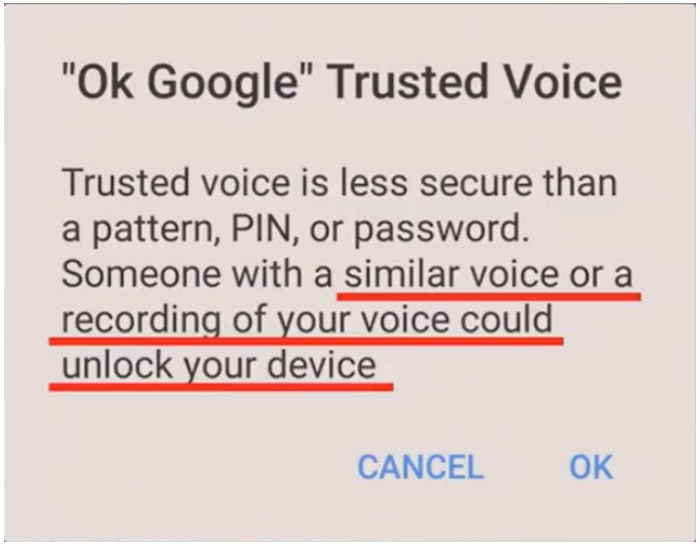
- A digital assistant that uses voice recognition, speech synthesis, and natural language processing (NLP) to provide a service.



MONTCLAIR STATE UNIVERSITY

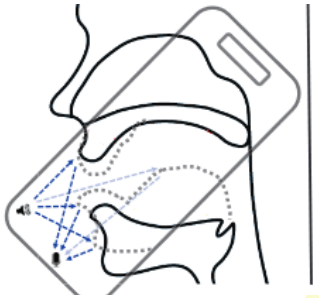
Threats of Voice

- Human voice is often exposed to the public
  - Attackers can “steal” or even generate victim’s voice
- Attackers can remotely replay stolen voices without physically being in the targeted smart home
  - Security issue → replay attack and mimicry attack



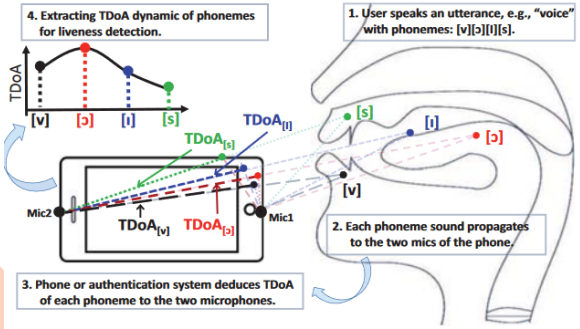
MONTCLAIR STATE UNIVERSITY

Previous Works – Liveness Detection

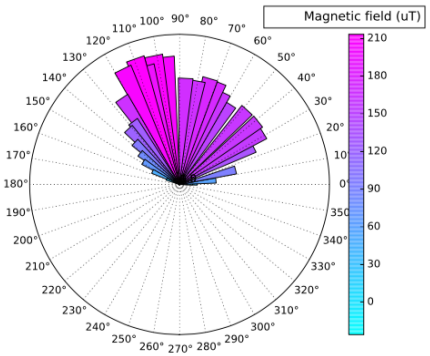


Lip motion based

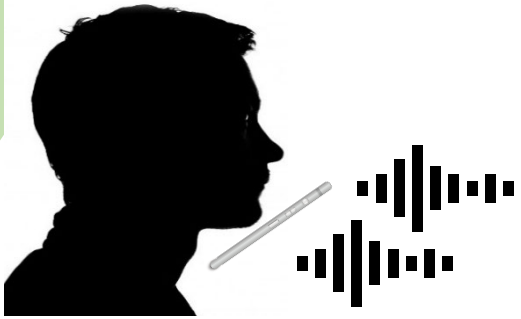
Phoneme location based



Magnetic fields of loudspeakers



Throat voice based



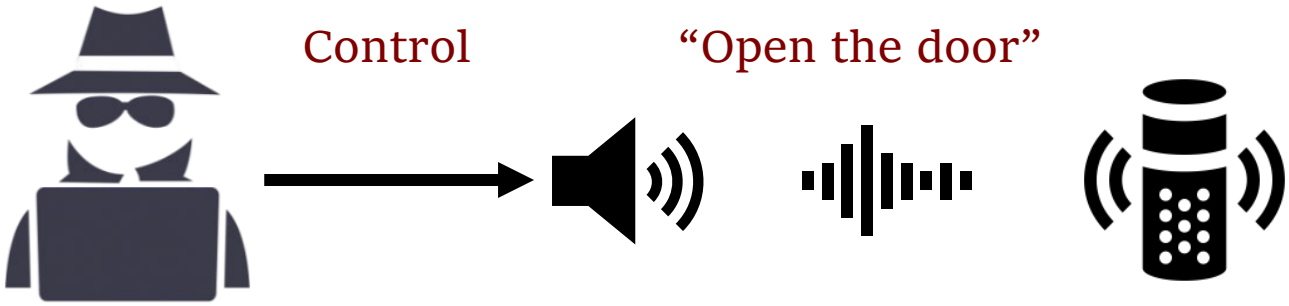
MONTCLAIR STATE UNIVERSITY

Goal

- Existing solutions cannot be used to protect the voice assistants on smart speakers
  - Short operation range
  - Need assistance from the user
- Goal: Design a new liveness detection system for smart speakers
  - User can use it anywhere in the smart environment
  - Low cost
  - Do not need assistance from the user

Attack Model

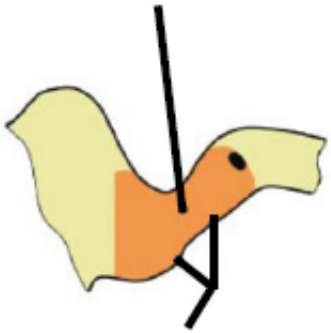
- Attackers can have control on some replay devices in the victim's smart home
  - E.g. speakers of smart TV
- Attackers replay malicious voice commands to control the victim smart home



Feasibility Study

- Our solution
  - Leverage the air pressure in ear canal
  - Mouth movement (opening and closing) can generate significant impact on the air pressure in ear canal
  - Collect air pressure using a tiny sensor in the earphones

Earmold with satisfactory retention with mouth closed



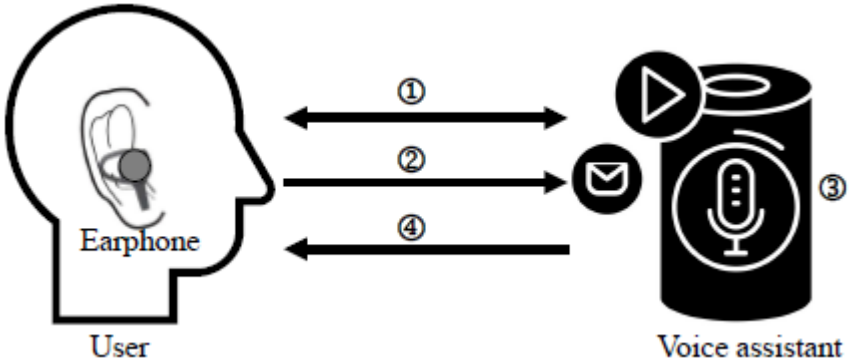
Ear canal shape with mouth closed

Earmold lacks retention with mouth opened

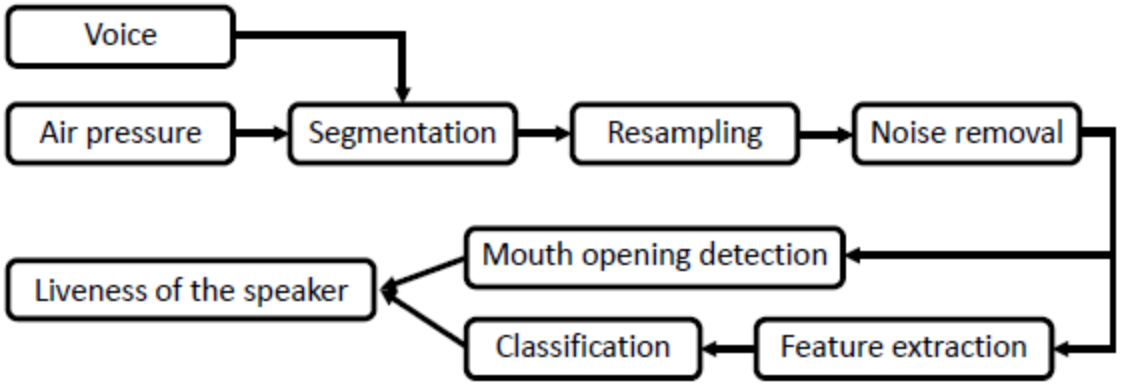
Ear canal shape with mouth opened

MONTCLAIR STATE UNIVERSITY

System Architecture



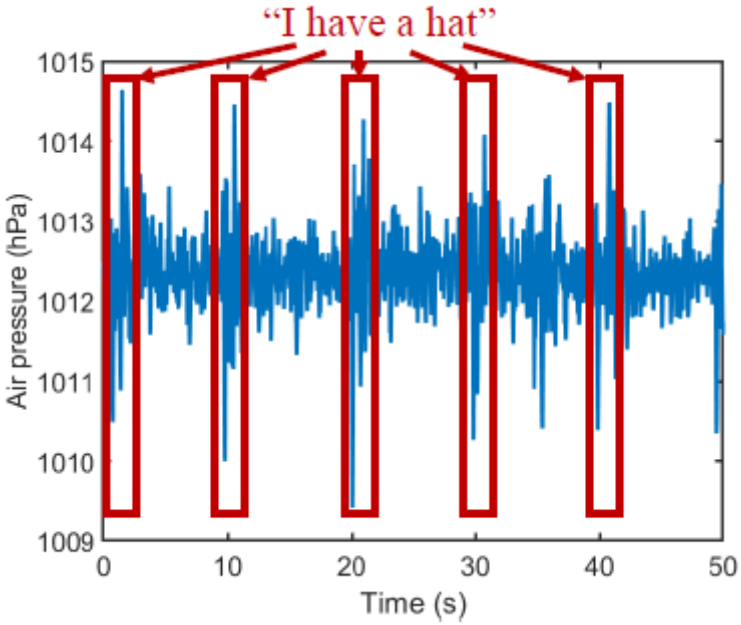
- Synchronization
- Send voices and air pressure
- Data processing
- Send the liveness detection results to the user



MONTCLAIR STATE UNIVERSITY

Feasibility Study

- Built a prototype to collect the ear canal pressure with a sampling rate of about 500 Hz and record the voice at the same time
- Ask a user to say a short sentence, “I have a hat”, every 10 seconds



MONTCLAIR STATE UNIVERSITY

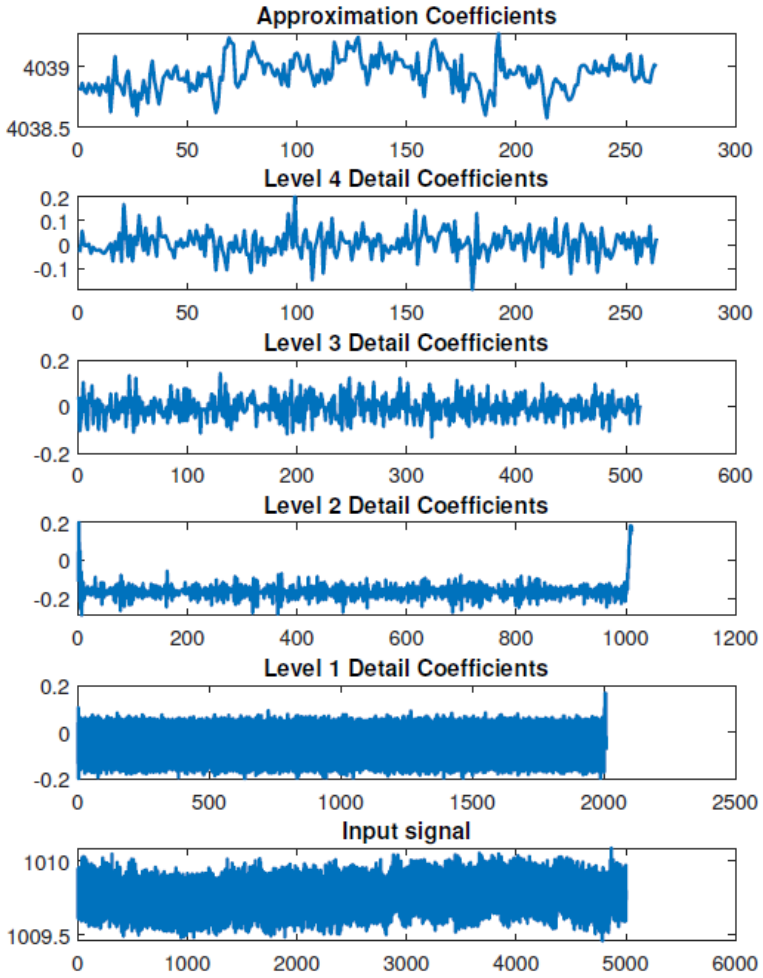
Preprocessing

- Signal segmentation
  - Get the segments of pressure signals that are influenced by the speeches
  - Hideen Markov Model-based word segmentation techniques

- Resampling
  - The raw pressure signal is not uniformly sampled
  - Filter the raw signal using a finite impulse response (FIR) filter
  - Normalize the result to account for the processing gain of the window and then change the sampling rate using a polyphase interpolation structure.

Preprocessing

- Noise removal
  - One-dimensional discrete wavelet decomposition with 4 levels
  - Leverage the approximation coefficients cA4 at the fourth level as the features



STATE UNIVERSITY

Mouth Opening Detection

- Calculate the short-term variance of the signal
  - Remove low-frequency noise
  - Short-term variances reach very high values when the user opens the mouth
  - Performing peak finding algorithm with a threshold

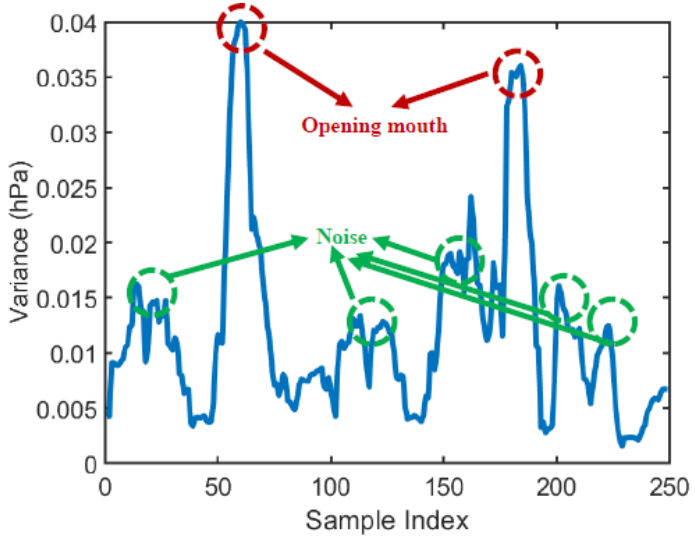


Fig. 10. Filtered variance signal.

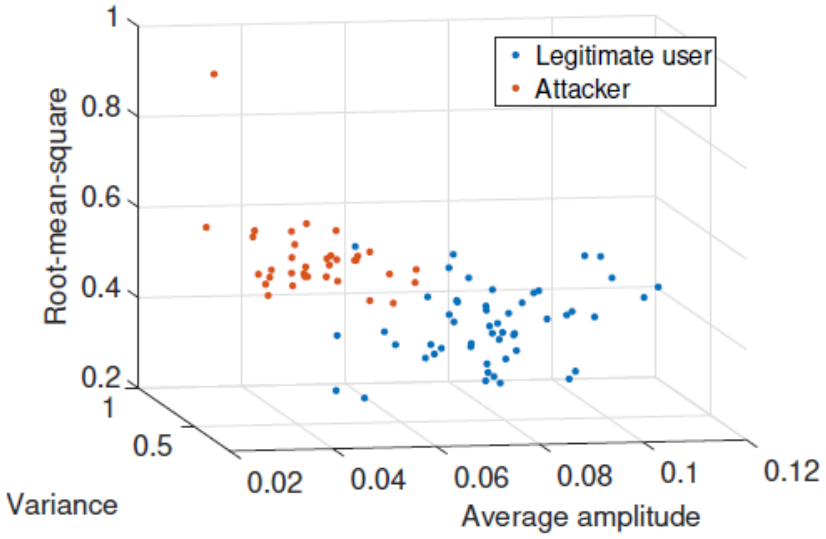
MONTCLAIR STATE UNIVERSITY

Enhanced Detection with Classification

- However, low-frequency noise may still exist in the filtered signal, which generates variances to the short-term variance signal
- An extra classifier to determine whether the short-term variance signal matches with those that are influenced by opening the mouth
- Two challenges in feature extraction
  - The absolute pressure values depend on the environment
  - The sampling rate of the variance signal is low (31.25Hz)

Enhanced Detection with Classification

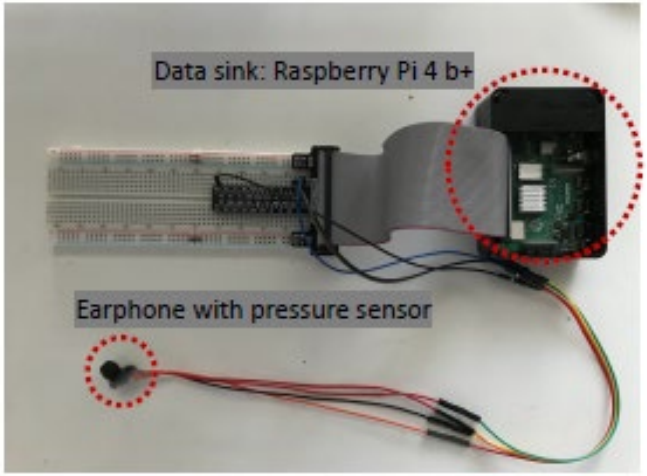
- To address these two challenges
  - Normalize the segmented short-term variance signal to a range (0,1]
  - Only extra features from the time-domain signal
    - Average amplitude, root-mean-square value, and overall variance
- Build the classifier using the Multiple Additive Regression Tree



MONTCLAIR STATE UNIVERSITY

Evaluation

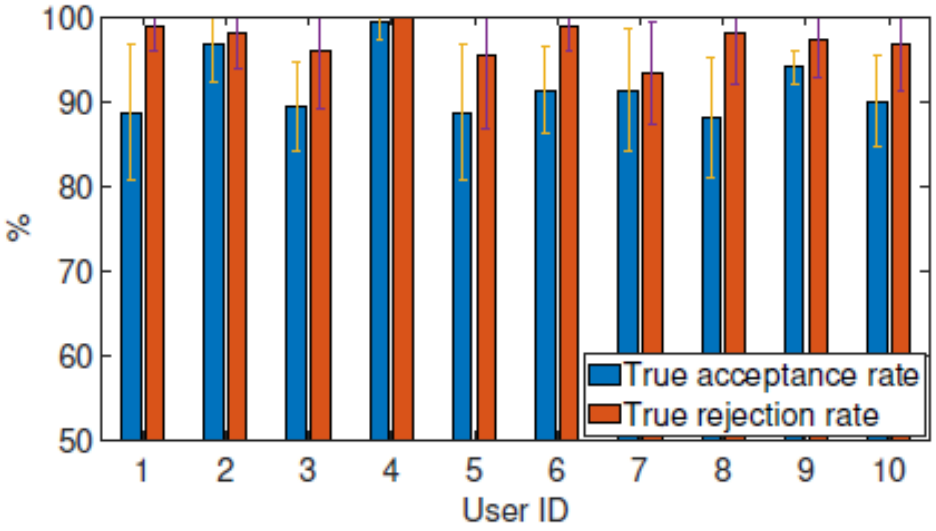
- Prototype
  - A pressure sensor (BMP 280)
  - A pair of earphones to hold the pressure sensor
  - A mini PC to collect the pressure sensor (Raspberry Pi)
  - A microphone to collect the voice
  - A data processing center
  - 10 volunteers involved



MONTCLAIR STATE UNIVERSITY

Overall Performance

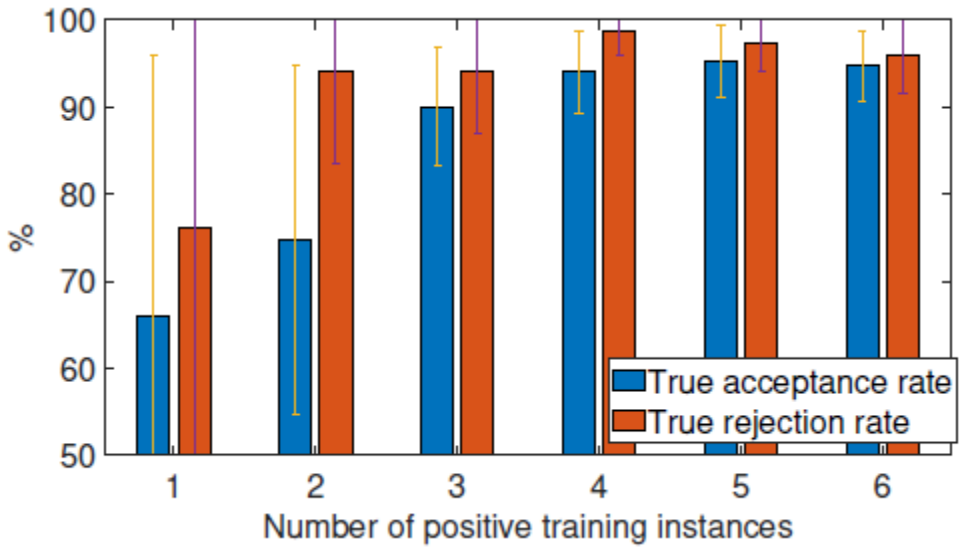
- Accept legitimate users with an average accuracy of about 91.72%
- Reject attackers with an average accuracy of 97.2%



MONTCLAIR STATE UNIVERSITY

Impact of the Number of Training Instances

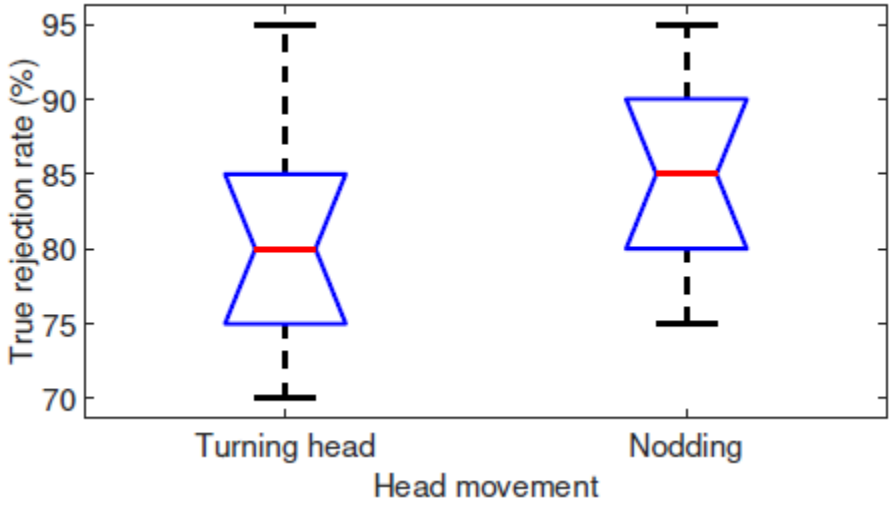
- Evaluate how many instances are needed from the new user



MONTCLAIR STATE UNIVERSITY

Impact of Facial Movements

- Some head movements (e.g. turning head) can also generate an impact on the air pressure in the ear canal.



MONTCLAIR STATE UNIVERSITY

Thank you

MONTCLAIR STATE UNIVERSITY