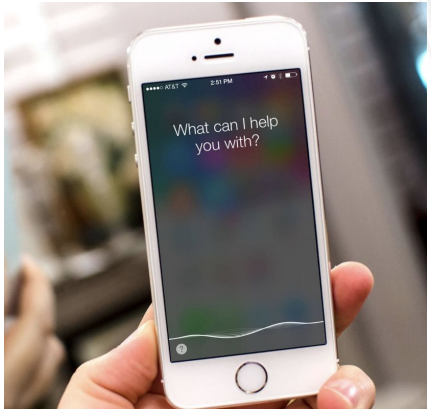
Continuous Authentication for Voice Assistants

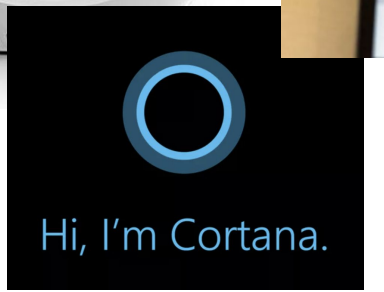
Huan Feng*, Kassem Fawaz*, and Kang G. Shin
Presented by Anousheh and Omer

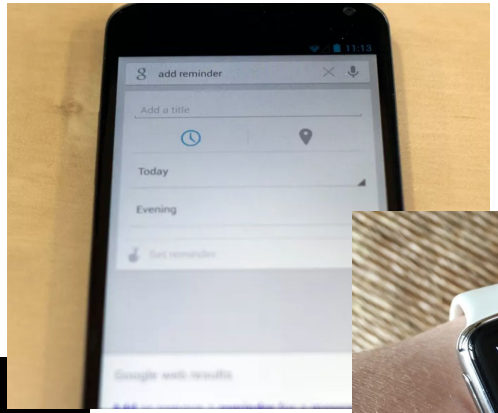
Overview

- Introduction/Existing Solutions and Novelty
- Human Speech Model
- System and Threat Models
- VAuth
- Matching Algorithm
- Phonetic-level Analysis
- Evaluation
- Discussion and Conclusion

Why voice user interface?







Introduction

- Voice as an User Interaction (UI) channel
  - Wearables, smart vehicles, home automation systems
- Security problem: open nature of the voice channel
  - Reply attacks, noise, impersonation
- **VAuth** is the first system providing continuous authentication for voice assistants
  - Adopted in wearables like eyeglasses, earphones/buds, necklaces
  - Match the body-surface vibrations and the microphone received speech signal

Existing solutions

Smartphone Voice Assistants

- **AuDroid**: a security mechanism that tracks the creation of audio communication channels explicitly and controls the information flows over these channels to prevent several types of attacks
  - requiring manual review for each potential voice command

Voice Authentication

- **Voice biometric**
  - rigorous training to perform well
  - no theoretical guarantee that they provide good security in general.
  - replay attacks.

Existing solutions(Cont'd)

Mobile Sensing

- It has been shown possible to infer keyboard strokes, smartphone touch inputs or passwords from acceleration information
- Most applications utilizing the correlation between sound and vibrations for health monitoring purposes, not continuous voice assistant security

Novelty

- **Continuous authentication**
  - Assumption of most authentication mechanisms (passwords, PINs, pattern, fingerprints) : the user has exclusive control of the device after authentication, not valid for voice assistants
  - **VAuth** provides ongoing speaker authentication
- **Improved security features**
  - Automated speech synthesis engines can construct a model of the owner's voice using very limited number of his/her voice samples
  - User has to unpair when losing **VAuth** token
- **Usability**
  - No user-specific training, immune to voice changes over time and different situations (where voice biometric approaches fail)

Human Speech Model

Source-filter Model

Human speech production has two processes:

- Voice source: vibration of vocal folds
- Filter: determined by resonant properties of vocal tracts including the effects of lips and tongue

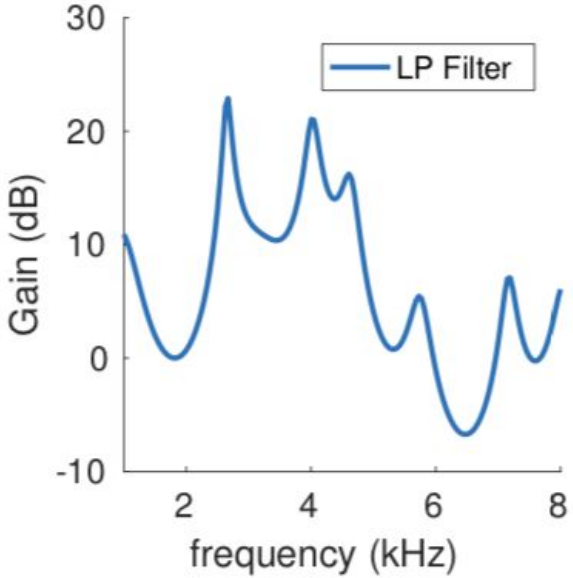


Fig. 2. Filter example of the vowel {i:}

Source-filter Model(Cont'd)

- **Glottal cycle length**: length of each glottal pulse (cycle)
- **Instantaneous fundamental frequency** (f0): inverse of glottal cycle length
- 80 Hz < f0 < 333Hz for human
- 0.003 sec < glottal cycle length < 0.0125 s
- Important feature of speaker recognition: the pitch changes pronouncing different phonemes

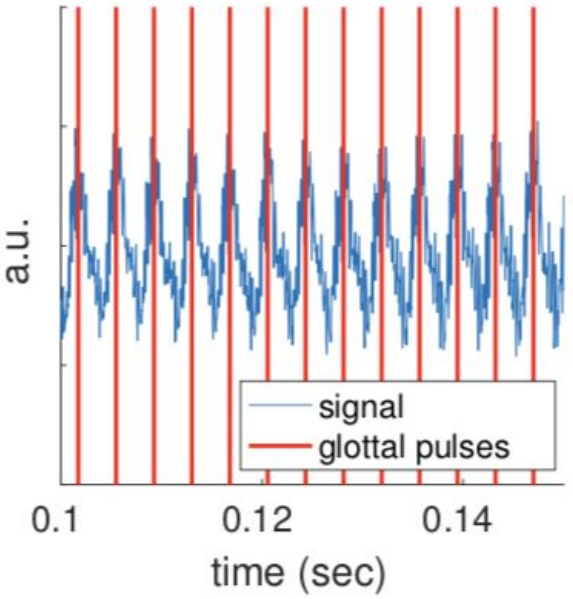
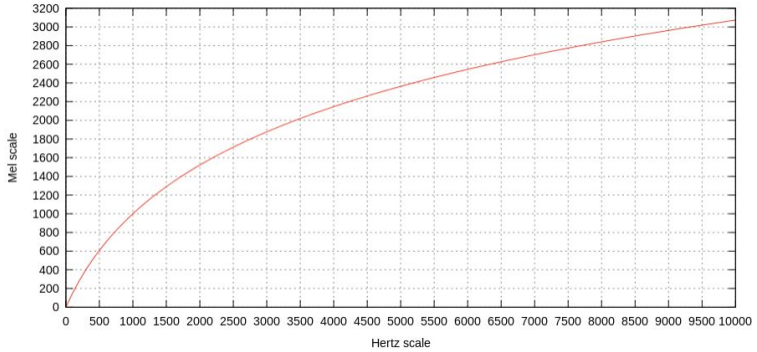


Fig 3. Voice source output

Speech Recognition and MFCC



Mel-frequency cepstral coefficients (MFCC):

- Most widely used feature for speech recognition
- Representation of the short-term power spectrum of a sound
- Steps:
  - Compute short-term Fourier transform
  - Scale the frequency axis to the non-linear Mel scale
  - Compute Discrete Cosine Transform(DCT) on the log of the power spectrum of each Mel band
- Works well in speech recognition, because it tracks the invariant feature of human speech across different users, but it can be attacked by generating voice segments with the same MFCC feature

System and Threat Models

VAuth System Model

VAuth components:

- **Wearable** : Housing an accelerometer touching user's skin at facial, throat, and sternum
- **Extended voice assistant** : Correlates accelerometer and microphone signal signals

Assumptions:

- Communication between two components is encrypted
- Wearable device serves as a secure token

Threat Model

The attacker wants to steal private information or conduct unauthorized operations by exploiting the voice assistant

- **Stealthy Attacks**
  - Injecting inaudible or incomprehensible voice commands through wireless signals or mangles voice commands
- **Biometric-override Attack**
  - Injecting voice commands by replying or impersonating victim's voice
  - Example: Google Now trusted voice feature is bypassed within five trials
- **Acoustic Injection Attack**
  - Generating a voice that has direct effect on the accelerometer like very loud music consisting embedded patterns of voice commands

VAuth

VAuth High-level Design

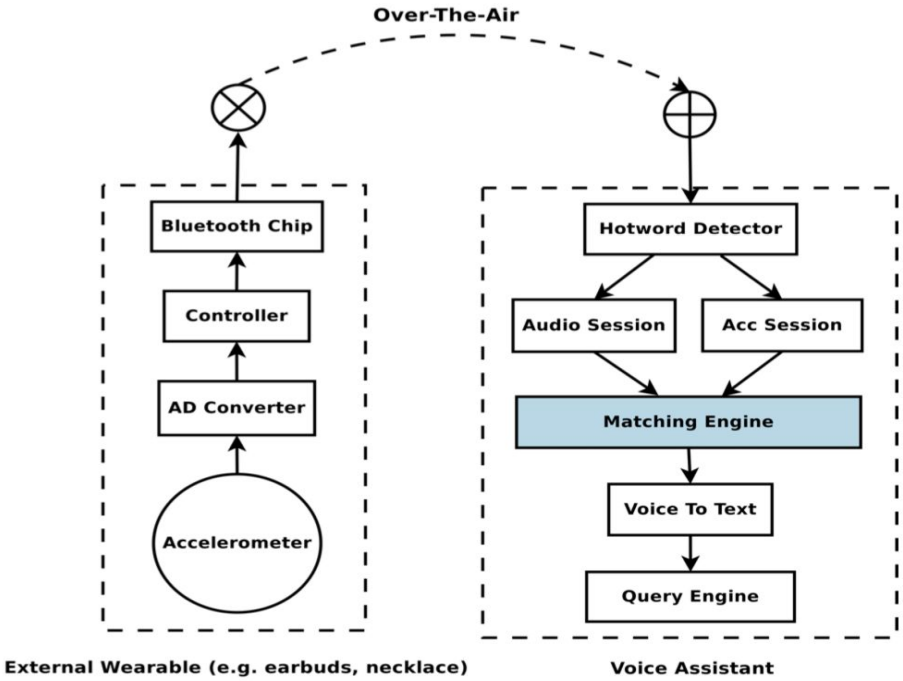


Fig 3. VAuth design components

Prototype

- Knowles BU-27135 miniature accelerometer with dimensions of 7.92*5.59*2.28 mm
- Accelerometer uses only z-axis and has bandwidth of 11KHz
- The system is integrated with Google Now voice assistant
- The microphone and accelerometer signals are sent to a Matlab-based sever performing matching and sending result to the voice assistant
- VAuth Intercepts both HotwordDetector and QueryEngine to establish required control flow

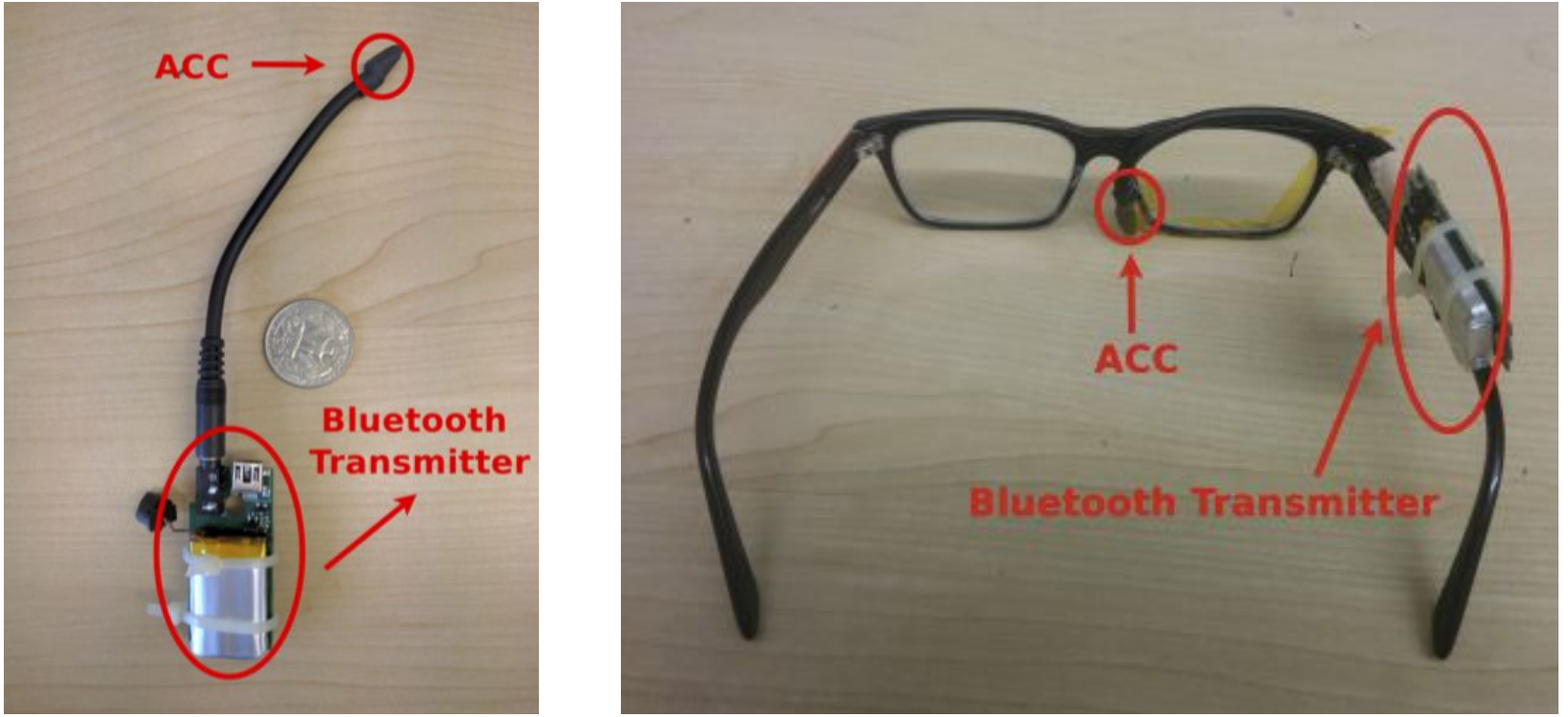


Fig. 1. Proposed prototype of VAuth

Usability

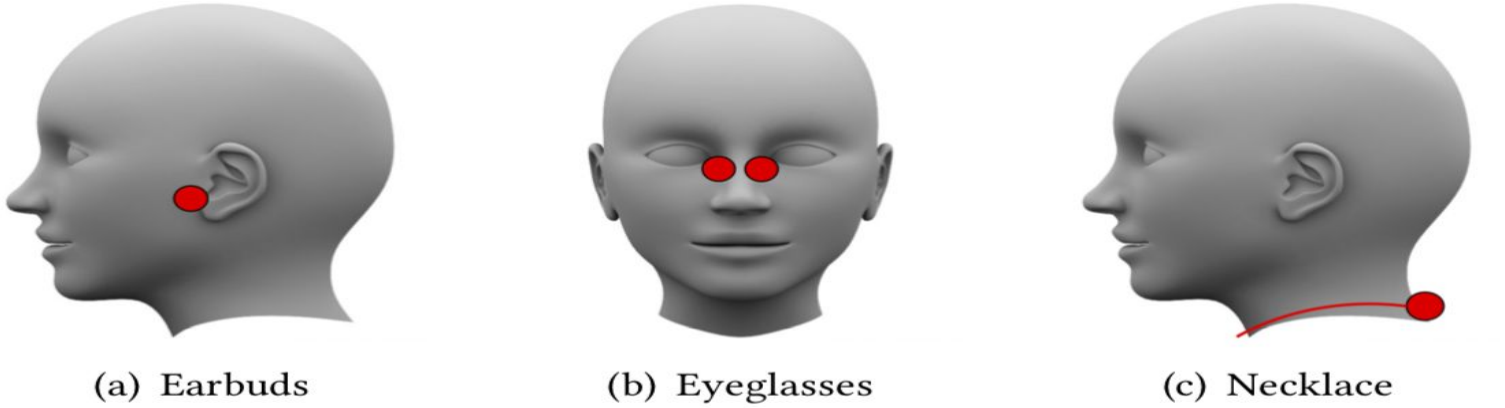


Fig 4. Wearable scenarios supported by VAuth

Usability Survey

- 952 participants, with experience using voice assistants,
  - 58% reported using a voice assistant at least once a week
- Questionnaire
  - USE questionnaire methodology
  - 7-point Likert scale(ranging from strongly agree to strongly disagree)

Fig. 5. A breakdown of respondent's wearability preference

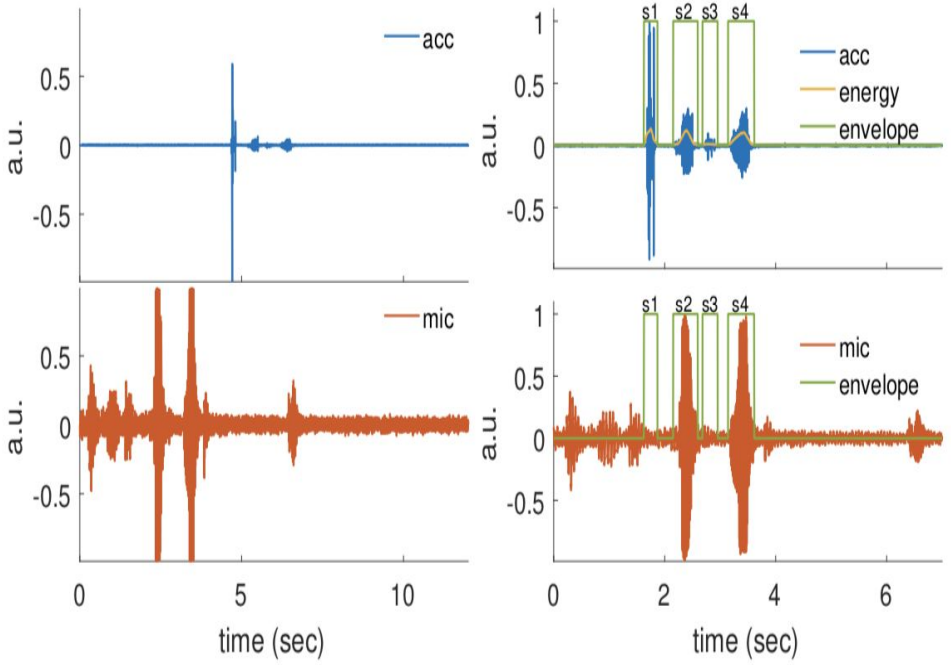
Matching Algorithm

Matching Algorithm Overview

- Inputs: speech and vibration signals and their sampling rate
- Output: decision value and a “cleaned” speech signal in case of match
- **Matching** algorithm stages:
  - Pre-processing
  - Speech segments analysis
  - Matching decision
- Running example
  - “cup” and “luck” words with a short pause between
  - 64 KHz and 44.1 KHz sampling frequency of speech and microphone signals

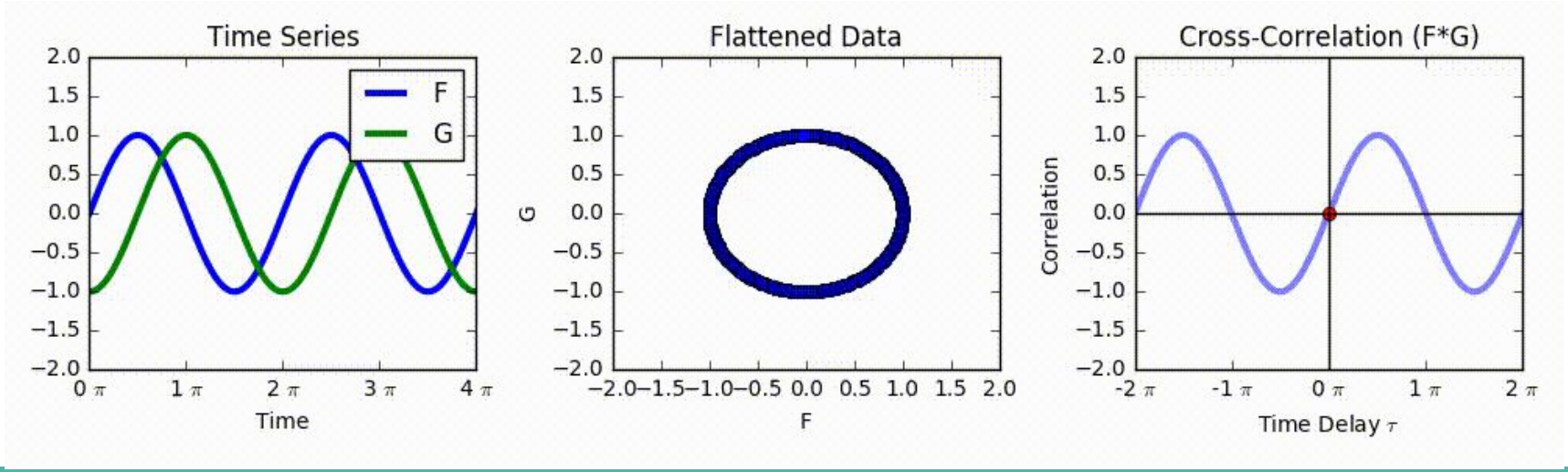
Pre-processing

- Highpass filter (Cut-off: 100Hz)
- Re-sampling acc and mic signals
- Normalization
- Aligning both signals to maximize their cross correlation
- Finding energy envelope of the accelerometer signal (High SNR)
- Applying accelerometer envelope to mic signal



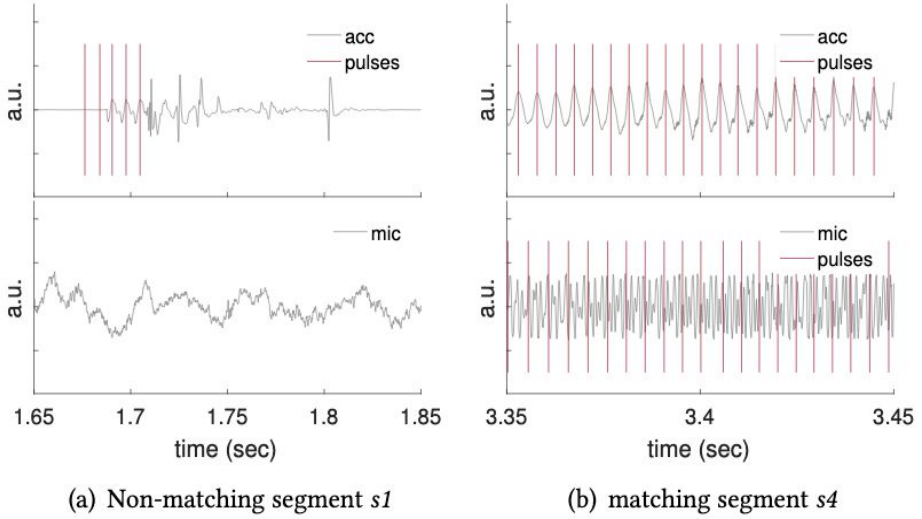
Cross correlation?

- Elementwise multiply two signals, and add the products.
- Normalized?
  - First normalize the signals to have the same range, then do the element wise multiplication.



Per-segment analysis

- Compare high energy segments to each other
- Find matching glottal cycles in the both data
- Freq must be within human range
- Relative pulse seq distance should be the same between the two
- Run normalized cross correlation between segments
- Delete the segment if any of these do not hold
- Keep if maximum correlation coefficient is within [-.25, .25]



(a) Non-matching segment *s1* (b) matching segment *s4*

Figure 7: Per-segment analysis stage of VAuth.

Matching decision

- Take “surviving” segments
- Run normalized cross correlation on the “surviving” segments as a whole.
- Use an SVM to map the result of the cross correlation to the *matching* or *non-matching* of the signals.

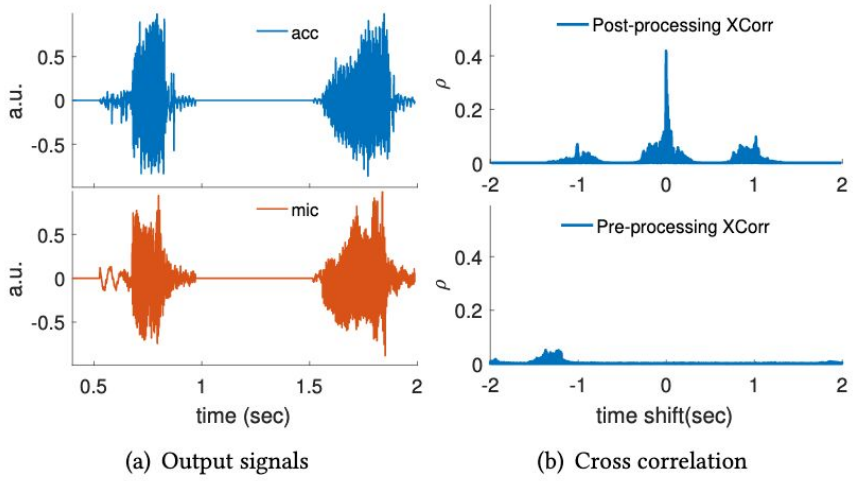


Figure 8: Matching decision stage of VAuth’s matching.

SVM details

- Feature set: take the max value of the Xcorr and sample 500 points to the right and 500 to the left of the max value. This gives a 1001 element vector.
- Classifier: Train SVM with Sequential Minimal Optimization algorithm. SVM has a polynomial kernel with degree 1.
- Training set: is the feature vectors labeled accordingly. They obtain this by generating every combination of microphone phoneme vs accelerometer phoneme. The recordings are generated form two people pronouncing the phonemes (more on this later).

PHONETIC-LEVEL ANALYSIS

Phonetic-level analysis

- Phonemes: an english word or sentence, spoken by a human, is necessarily a combination of english phonemes.
- Essentially the fundamental sounds we make to speak.
- 44 of them in english.
- Recruit 2 people (male,female)
- Each participant records 2 examples for each phoneme.

Table 1: The IPA chart of English phonetics.

Vowel	Examples	Conso-nants	Examples
ʌ	CUP, LUCK	b	BAD, LAB
ɑː	ARM, FATHER	d	DID, LADY
æ	CAT, BLACK	f	FIND, IF
e	MET, BED	g	GIVE, FLAG
ə	AWAY, CINEMA	h	HOW, HELLO
ɜːʳ	TURN, LEARN	j	YES, YELLOW
ɪ	HIT, SITTING	k	CAT, BACK
iː	SEE, HEAT	l	LEG, LITTLE
ɒ	HOT, ROCK	m	MAN, LEMON
ɔː	CALL, FOUR	n	NO, TEN
ʊ	PUT, COULD	ŋ	SING, FINGER
uː	BLUE, FOOD	p	PET, MAP
ɑɪ	FIVE, EYE	r	RED, TRY
ɑʊ	NOW, OUT	s	SUN, MISS
eɪ	SAY, EIGHT	ʃ	SHE, CRASH
oʊ	GO, HOME	t	TEA, GETTING
ɔɪ	BOY, JOIN	ʧ	CHECK, CHURCH
eəʳ	WHERE, AIR	θ	THINK, BOTH
ɪəʳ	NEAR, HERE	ð	THIS, MOTHER
ʊəʳ	PURE, TOURIST	v	VOICE, FIVE
-	-	w	WET, WINDOW
-	-	z	ZOO, LAZY
-	-	ʒ	PLEASURE, VISION
-	-	ʤ	JUST, LARGE

Phonetic-level analysis cont.

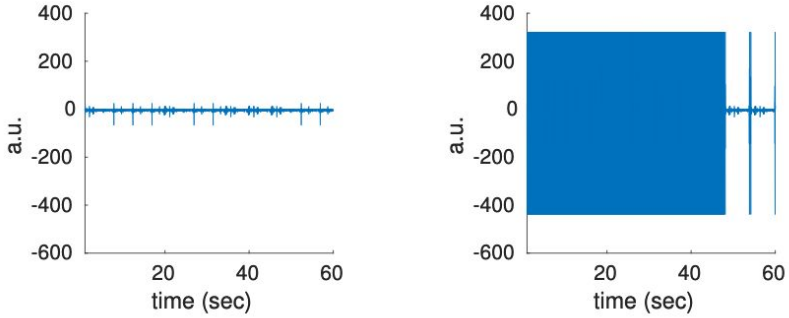
- Idea: Why not just use the accelerometer data and do Automatic Speaker Recognition?
  - All phonemes register vibrations on the accelerometer.
  - Use “state-of-the-art” Nuance Automatic Speaker Recognition.
- Doesn’t work, the accelerometer samples are too low fidelity.

Phonetic-level analysis cont.

- Phonemes detection accuracy?
  - 176 samples in total (2 speaker, 2 examples per phoneme)
- What happens when there is voice but not from the user?
  - No false positives in their tests.
  - Doesn't necessarily mean there isn't an attack vector here.

Table 2: The detection accuracy for the English phonemes.

microphone	accelerometer	TP (%)	FP (%)
consonants	consonants	90	0.2
consonants	vowels	-	1.0
vowels	consonants	-	0.2
vowels	vowels	100	1.7
all	all	94	0.7



(a) Low energy and white noise

(b) High energy and periodic noise

Figure 9: Examples of tested noise signals.

EVALUATION

Evaluation

- Test the system for a number of different users.
- 95% accuracy (TPs)
- Doesn’t work for Korean.
- Evaluate different security scenarios
- Evaluate the delay and energy problems

User study

- IRB approval ✓
  - What about the previous stuff?
- 18 users
  - Recruitment?
  - Demographics?
- 3 positions of the device
- 2 user states: jogging and still
- 30 phrases
- Each user do the 6 combinations
- Voice assistant is Google Now.

Table 3: The list of commands to evaluate VAuth.

Command	Command
1. How old is Neil deGrasse Tyson?	**16.** Remind me to buy coffee at 7am from Starbucks
2. What does colloquial mean?	**17.** What is my schedule for tomorrow?
3. What time is it now in Tokyo?	**18.** Where's my Amazon package?
4. Search for professional photography tips	**19.** Make a note: update my router firmware
5. Show me pictures of the Leaning Tower of Pisa	**20.** Find Florence Ion's phone number
6. Do I need an umbrella today? What's the weather like?	**21.** Show me my bills due this week
7. What is the Google stock price?	**22.** Show me my last messages.
8. What's 135 divided by 7.5?	**23.** Call Jon Smith on speakerphone
9. Search Tumblr for cat pictures	**24.** Text Susie great job on that feature yesterday
10. Open greenbot.com	**25.** Where is the nearest sushi restaurant?
11. Take a picture	**26.** Show me restaurants near my hotel
12. Open Spotify	**27.** Play some music
13. Turn on Bluetooth	**28.** What's this song?
14. What's the tip for 123 dollars?	**29.** Did the Giants win today?
15. Set an alarm for 6:30 am	**30.** How do you say good night in Japanese?

User study

- Still: 97% TPs, 0.09% FPs
  - 2 outliers, low volume

- Jogging: ?
  - Outliers situation seems to be better
  - People might be speaking louder because they are jogging.

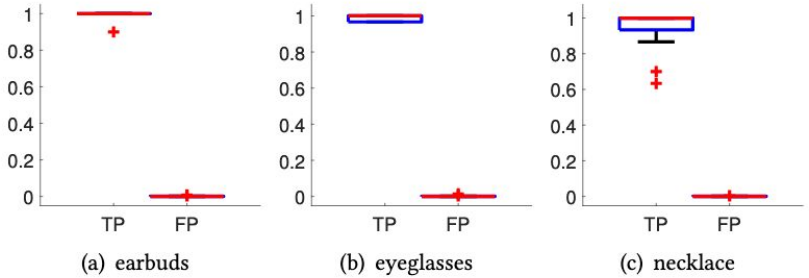


Figure 10: The detection accuracy of VAuth for the 18 users in the still position.

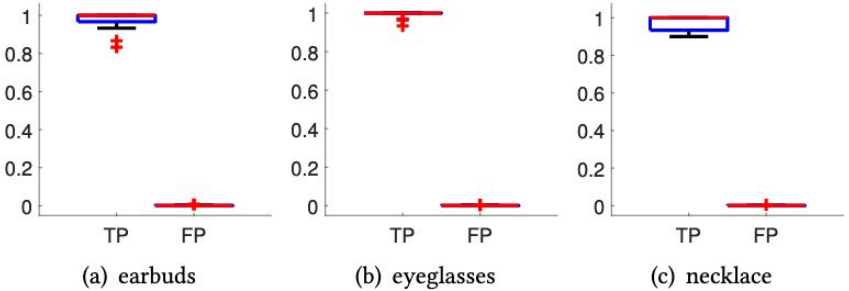


Figure 12: The detection accuracy of VAuth for the 18 users in the moving position.

User study

- Different languages?
- Recruit 4 new participants
  - Arabic
  - Chinese
  - Korean
  - Persian
- Works surprisingly well (97% TPs)
  - Korean lacks nasal sounds

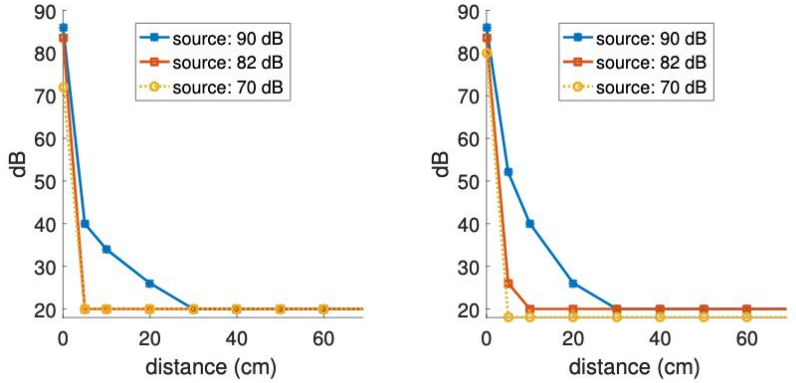
Table 4: The detection accuracy of VAuth for the 4 different languages.

Scenario	Language	TP (%)	FP (%)
earbuds	Arabic	100	0.1
	Chinese	100	0
	Korean	100	0
	Persian	96.7	0.1
eyeglasses	Arabic	100	0
	Chinese	96.7	0
	Korean	76.7	0
	Persian	96.7	0
necklace	Arabic	100	0
	Chinese	96.7	0
	Korean	96.7	0
	Persian	100	0

Security

- Silent user:
  - Completely prevents the **stealthy** and **biometric override** attacker.
  - The **Acoustic Injector** cannot make the accelerometer register stuff beyond a cutoff.

Sce-nario	Adversary	Example	Silent User	Speaking User
A	Stealthy	mangled voice, wireless-based	✔	✔
B	Biometric Override	replay, user impersonation	✔	✔
C	Acoustic Injection	direct communication, loud voice	distance cut-off	distance cut-off



(a) Exposed accelerometer (b) Covered accelerometer

Figure 13: The magnitude of the sensed over-the-air vibrations by the accelerometer as a function of the distance between the sound source and the accelerometer.

Security

- Speaking user:
  - **Stealthy** attacker: create the MFCC representation of the spoken words, construct a new command that has the same MFCC and send the new command to VAuth. Doesn't work, the acceleration and mic data don't match up even though the mic data for the user and attacker do.
  - **Biometric override** and **acoustic injection** fail similarly to the silent user.

Sce-nario	Adversary	Example	Silent User	Speaking User
A	Stealthy	mangled voice, wireless-based	✔	✔
B	Biometric Override	replay, user impersonation	✔	✔
C	Acoustic Injection	direct communication, loud voice	distance cut-off	distance cut-off

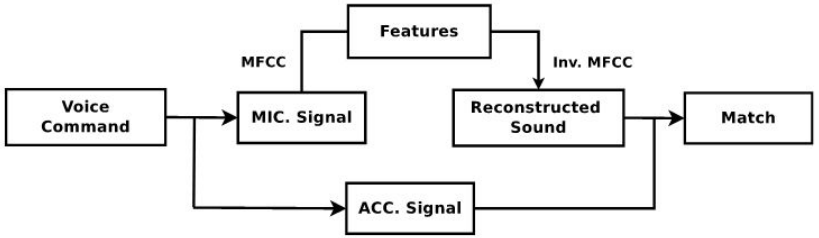


Figure 14: The flow of the mangled voice analysis.

Delay and Energy

- Delay:
  - 300-830ms, μ: 364ms when match is successful.
  - 230-760ms, μ: 319ms when match unsuccessful.
  - < 1 second for 30 word sentences.
  - Could be optimized further with a server implementation.
- Energy:
  - Mostly sits idle.
  - 100 voice commands per day with 500mAh battery should last a week.
  - If integrated into another wearable, only introduces accelerometer overhead.

DISCUSSION & CONCLUSION

Discussion & Conclusion

- The system requires new hardware.
  - This could be engineered into existing wearables.
- The system has energy constraints
- Uses accelerometer as opposed to microphones. Microphones are more vulnerable towards attacks.