Active Voice Authentication

Qi (Peter) Li, Ph.D.
Li Creative Technologies, Inc.

Fred B.H. Juang, Ph.D.
Georgia Tech

October 29, 2014

Team Members

- Qi (Peter) Li, Ph.D., PI, LcT, Florham Park, NJ
- B.-H. (Fred) Juang, Ph.D. Co-PI, Georgia Tech, Atlanta, GA
- Xuling Lu, Ph.D., Researcher, LcT, Florham Park, NJ
- Manli Zhu, Ph.D., Researcher, LcT, Florham Park, NJ
- Craig Adams, M.A., Technical Writer, LcT, Florham Park, NJ
- Umair Altaf, Ph.D. Student, Georgia Tech, Atlanta, GA
- Zhong Meng, Ph.D. Student, Georgia Tech, Atlanta, GA
- Chao Weng, Ph.D. Student, Georgia Tech, Atlanta, GA

Active Voice Authentication (AVA)

Two Operating Modes:

- **Unlock**: user utters a passphrase to gain access to the device
- **Monitor**: the device continuously monitors the user ID status whenever voice apps are in use

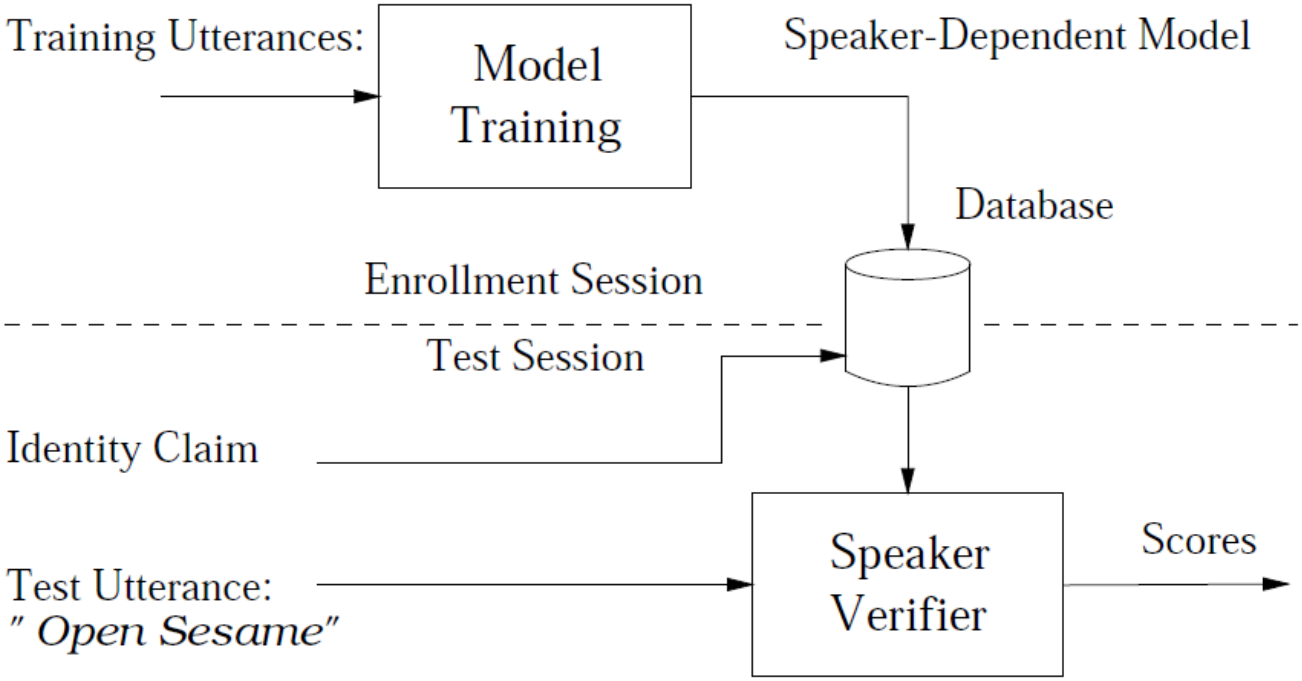
Supported by two Speaker Verification (SV) Technologies:

- **Unlock**: Text-Dependent SV, verifying both the spoken content (if it matches the passphrase) and the speaker's voice characteristics
- **Monitor**: Text-Independent SV, focusing on the speaker's voice characteristics only, regardless of the spoken content

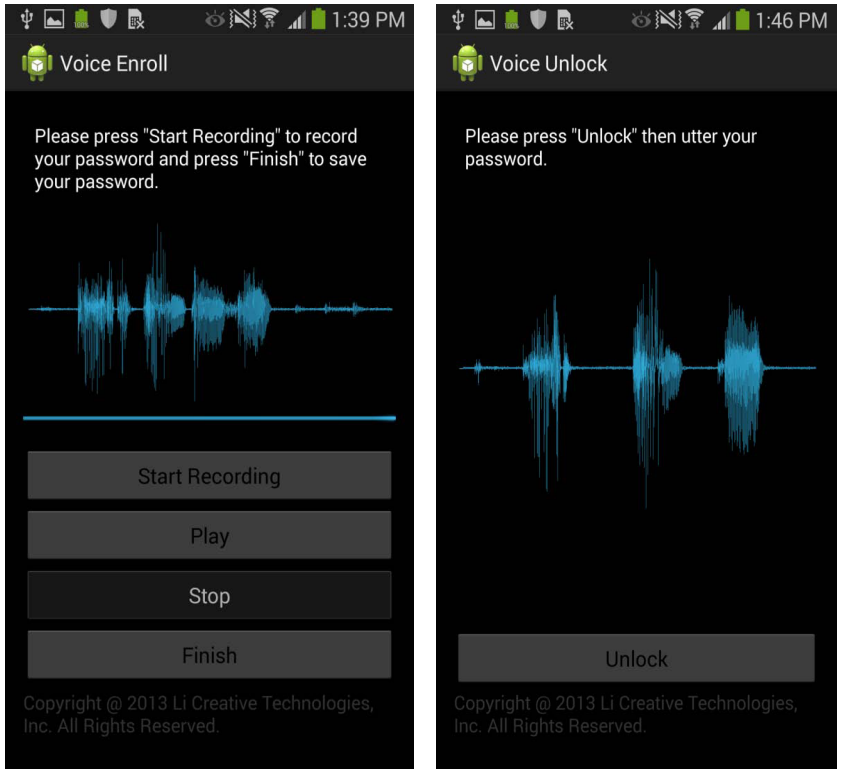
Active Voice Authentication - Unlock (AVA)

Li Creative Technologies, Inc.

Text-Dependent Speaker Verification

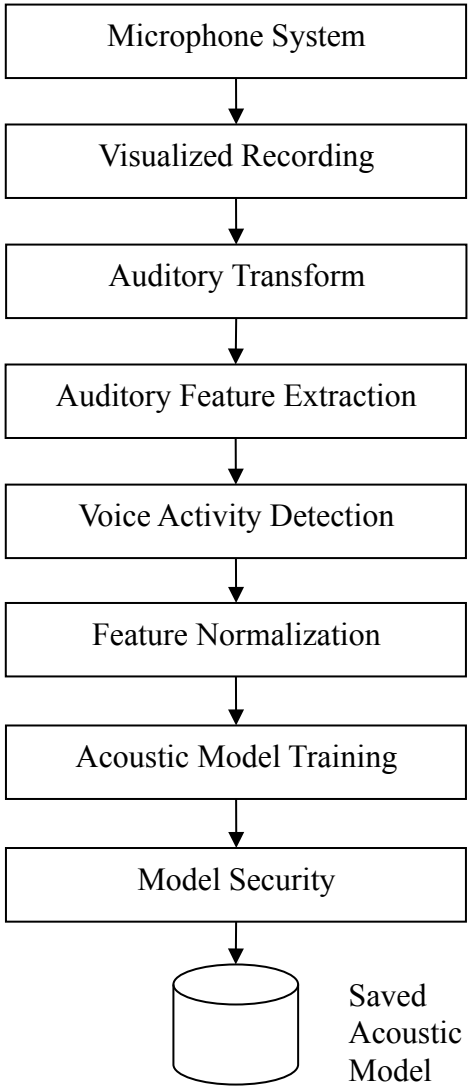


Unlock Software

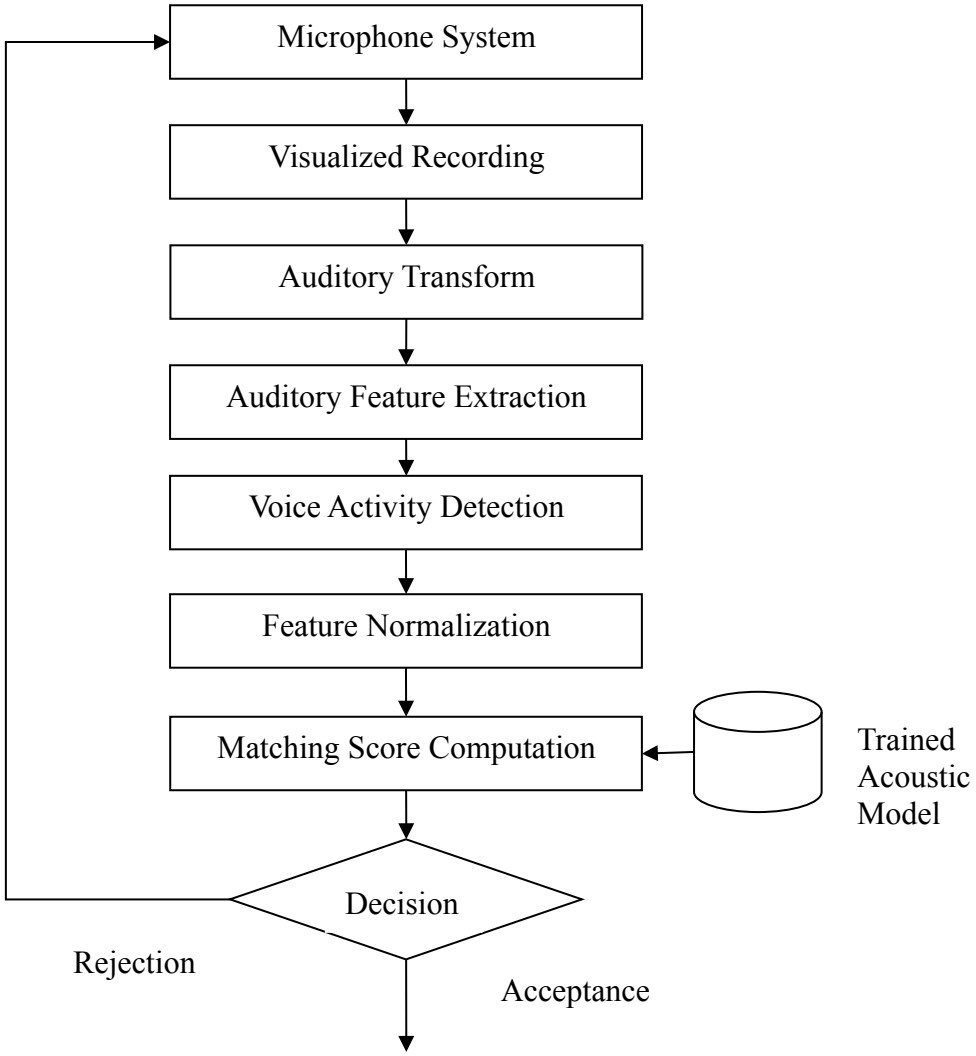


- Objective: Accept true speaker and reject imposters by voice passphrase
- Text-dependent speaker verification
- Training Mode:
  - The system prompts the user to utter a passphrase in 1 - 3 seconds
  - A speaker-dependent model is trained and saved in the device
- Testing mode:
  - Press button to talk
  - System responds immediately

Software Structure -- Enrollment

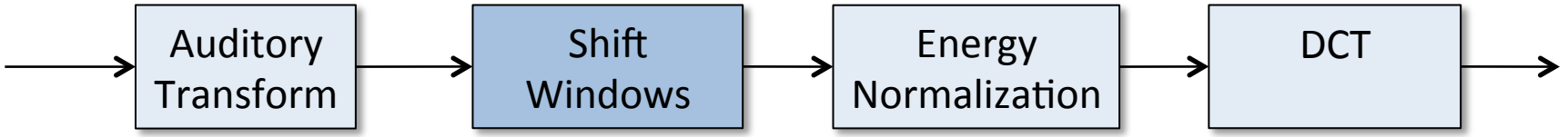


Software Structure -- Unlock

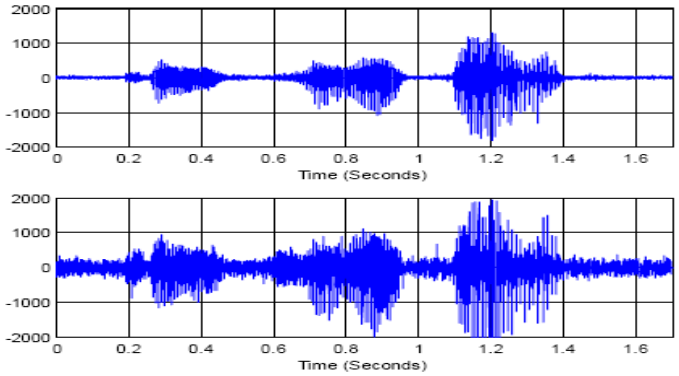


Feature Extraction

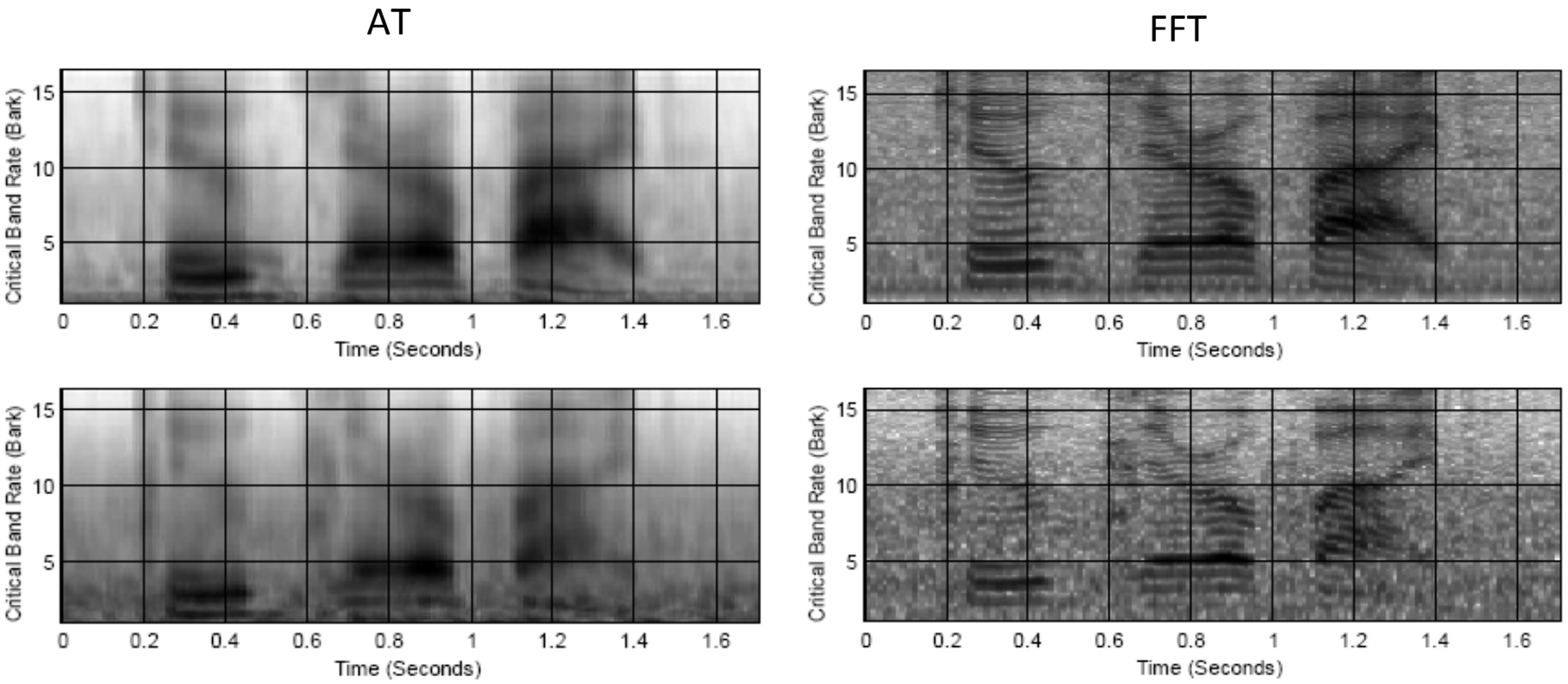
Speech
Waveform



AFCCs

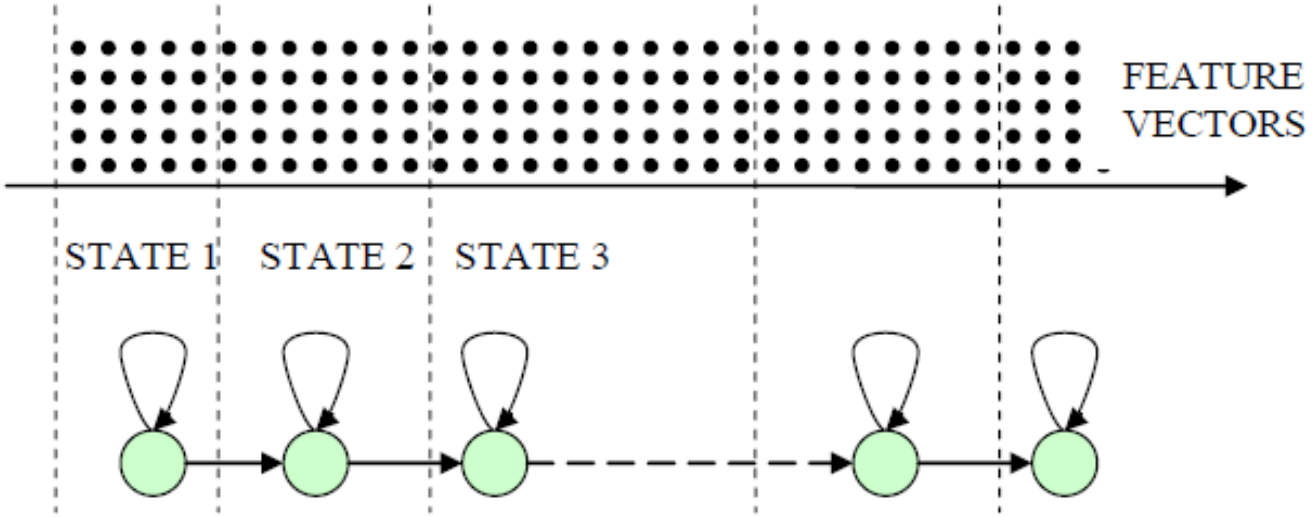


Auditory Transform vs. FFT Spectrograms



- Both spectrogram plotted on the Bark scale.
- FFT spectrograms have more noise and harmonics.
- AT spectrograms are more robust while keeping pitch and formants.

Hidden Markov Model



- Left-to-right HMM
- Each state is a GMM or a VQ model
- The model can be trained discriminatively to separate true speaker from imposters

Database Evaluation

- Randomly selected subset of 104 volunteers (56 females, 48 males)
- Each subject has 9 sessions
- Each session has 30 different passphrases in the same order
- Each session uses different recording handsets
- Normal acoustic environments

Situation 1

- All subjects use different passphrase
- This is close to real application scenario
- Similar to other biometric tests
- For each subject:
  - Training: One passphrase, 1 (3-4 seconds)
  - Testing
    - True speaker: 8 passphrases
      (the first passphrase from session 2 to session 9)
    - Imposters: 26,883 passphrases
      (103 imposters x 9 sessions x 29 passphrases)

Situation*	False Rejection	False Acceptance
All subjects use different passphrases	0.0%	0.00091%**

* Normal acoustic environments; ** One FA error in 100,000 trials.

Situation 2

- All subjects use the same fixed passphrase
- Worst case scenario
- This should not happen in real-world applications
- For each subject:
  - Training: One passphrase, the first passphrase from session 1 (3-4 seconds)
  - Testing
    - True speaker: 8 passphrases
      (the first passphrase from session 2 to session 9)
    - Imposters: 824 passphrases
      (103 imposters x 8 sessions x 1 passphrases)

Voice Biometrics	Average Equal Error Rate
Worst Case Scenario	2.61%

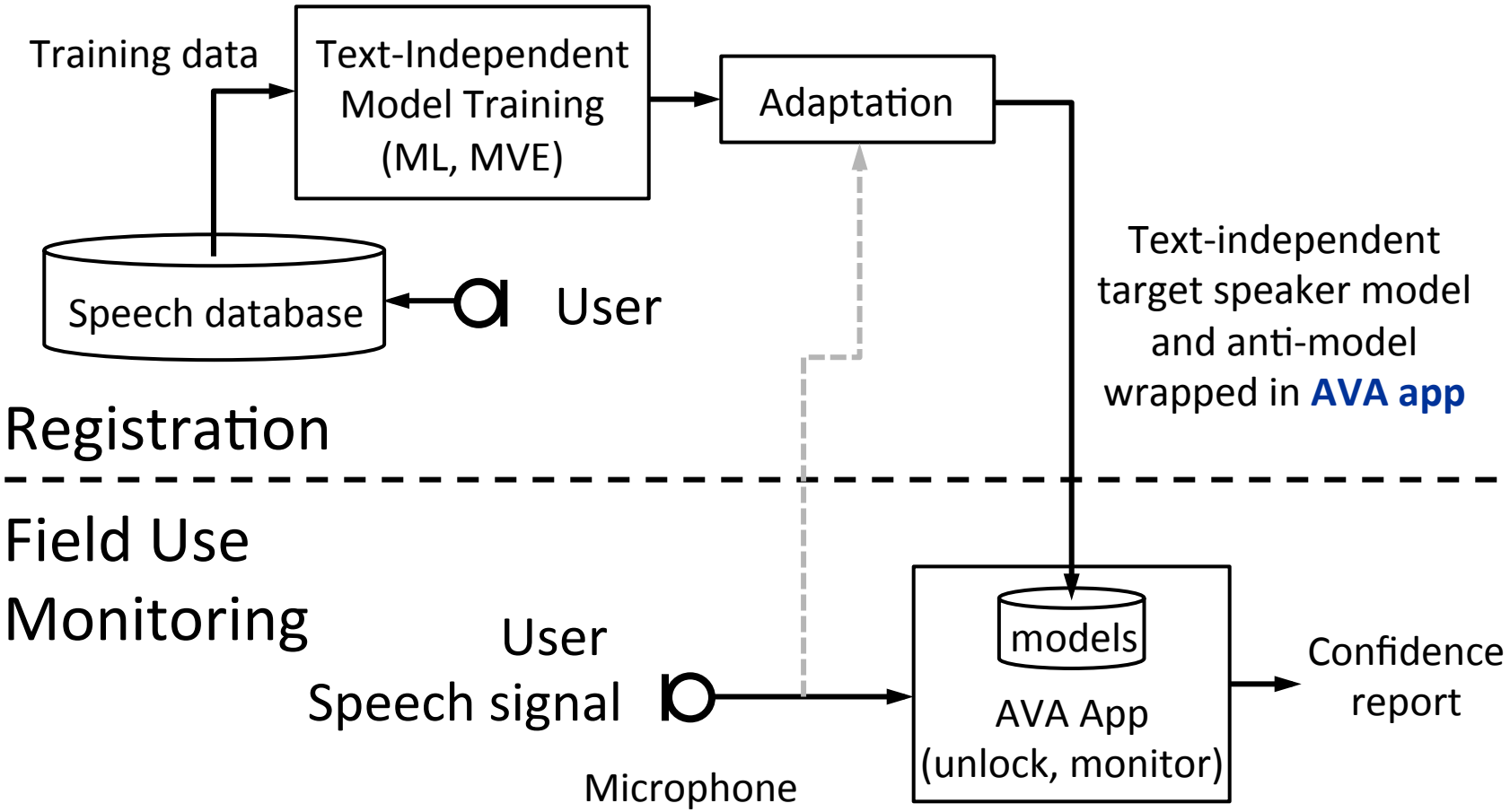
Unlock Summary

- Text-dependent speaker verification
- One passphrase for training
- Language independent
- Our new algorithm is simpler, faster, and more robust than traditional approaches
- Very low error rates

Active Voice Authentication - Monitor (AVA)

Georgia Institute of Technology

Operation Stages of Monitor Mode



Spoken Material for AVA Registration

Each talker speaks the following:

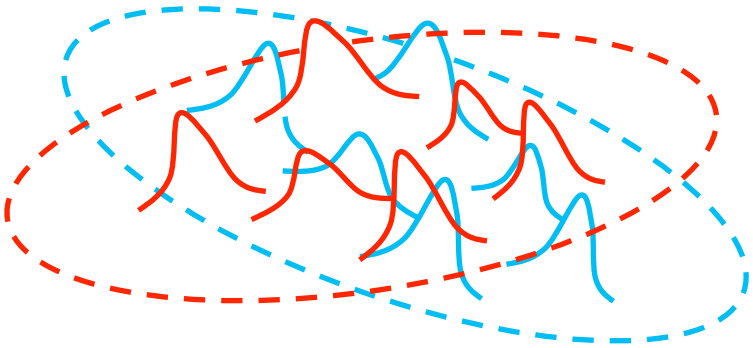
- **The Rainbow Passage**
- **A passphrase**
- **Phonetically Balanced Sentences** (20, randomly chosen)
- **Digit Pairs** (30, randomly chosen): e.g., “two eight”

❖ Finalization of the registration material (to shorten the process) will depend on system performance

❖ We have found even with 105 s (1’45”) training material, EER is still around 5-6% depending on model configuration. Performance detail follows.

Text-Independent Talker Model

- Based on characterization of the overall acoustic spectral space spanned by the target talker's speech
- Two essential models:
  - Partitioned spectral prototypes: via Vector Quantization (VQ)
  - Mixture densities: via Hidden Markov Model (HMM), possibly with sequential dependence



Model Estimation & Hypothesis-Testing

Model Training Criteria:

X = Talker Data; $\overline{X}$ = Imposter Data; Λ = Talker Model; $\overline{\Lambda}$ = Anti - Model

- Maximum Likelihood

$$\Lambda = \arg\max_{\Lambda'} P(X \mid \Lambda') \qquad \overline{\Lambda} = \arg\max_{\Lambda'} P(\overline{X} \mid \Lambda')$$

- Minimum Verification Error (MVE)

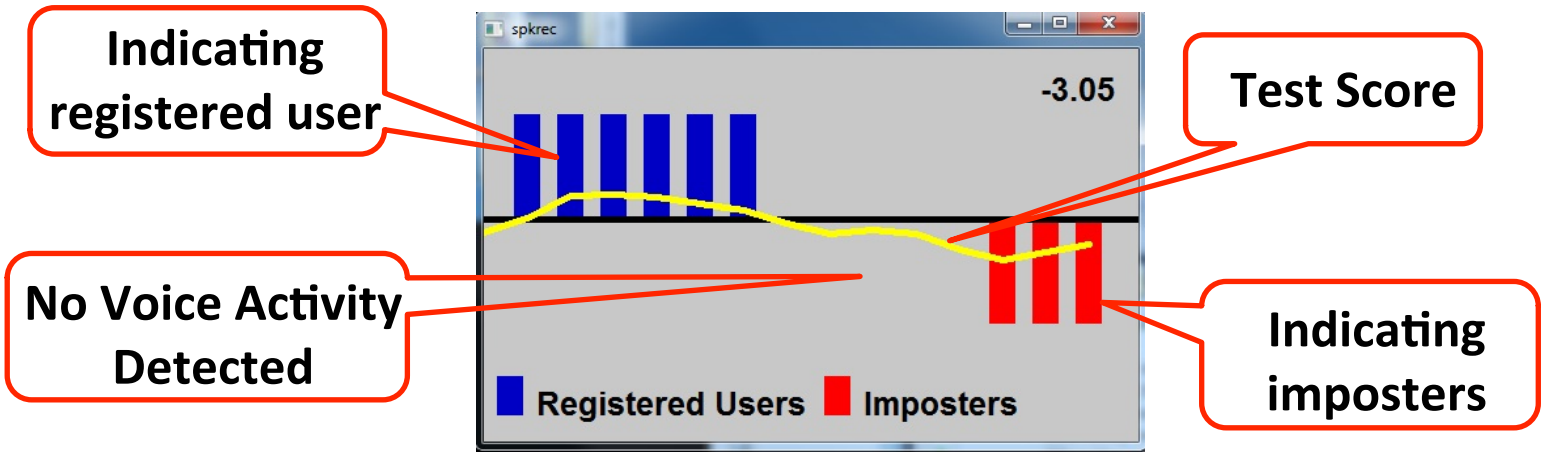
$$(\Lambda, \overline{\Lambda}) = \arg\min_{(\Lambda', \Lambda'')} \left[E\left\{ \sigma\left(\log \frac{P(X \mid \Lambda'')}{P(X \mid \Lambda')} \right) \right\} + \lambda E\left\{ \sigma\left(\log \frac{P(\overline{X} \mid \Lambda')}{P(\overline{X} \mid \Lambda'')} \right) \right\} \right]$$

Hypothesis-Testing (score reported at t_2):

$$\beta_{t_2} = \log \frac{P(\widetilde{X}_{t_1}^{t_2} \mid \Lambda)}{P(\widetilde{X}_{t_1}^{t_2} \mid \overline{\Lambda})} \quad \text{where } \widetilde{X}_{t_1}^{t_2} \text{ is unknown input in } (t_1, t_2) \text{ to be verified}$$

Monitor Mode – Sequential Tests

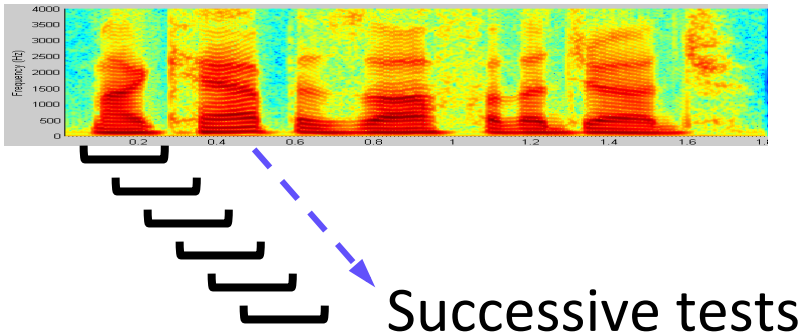
- The monitor module performs hypothesis testing on successive segments of speech data (currently set at 1 second)
- It reports test result (blue=confirm, red=deny) at a specified interval, modulated by confirmed voice activities



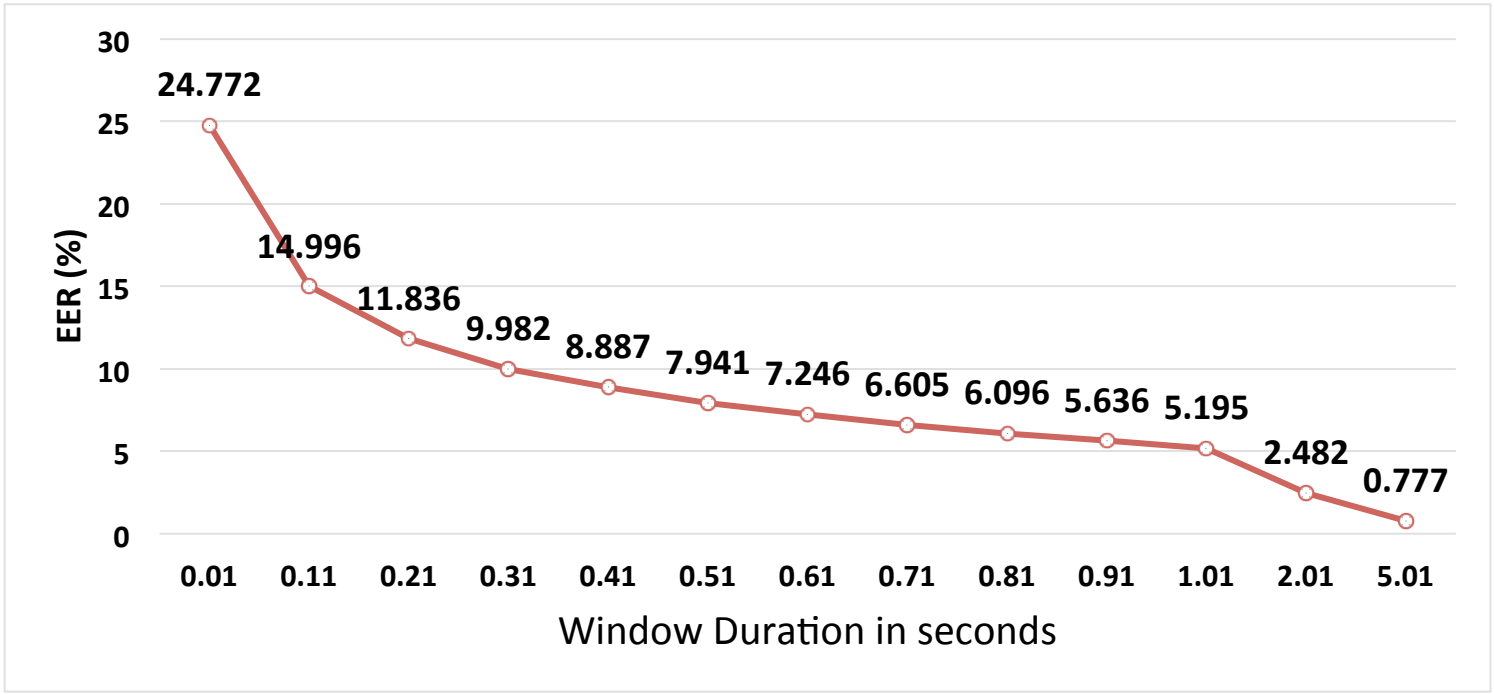
$$\beta_{t_1} = \log \frac{P(\tilde{X}^{t_1}_{t_1-\tau} \mid \Lambda)}{P(\tilde{X}^{t_1}_{t_1-\tau} \mid \overline{\Lambda})}$$

where $\tilde{X}^{t_1}_{t_1-\tau}$ is unknown input in $(t_1\text{-}\tau, t_1)$ to be verified

EER Performance vs. Test Window (Baseline)



- Successive tests performed on the microphone signal
- Each test spans over a window/segment of a certain duration



EER Evaluation on AVA Database (MAP + MVE with VAD)

Window Size (s)	Total Number of Mixtures	Number of States	Number of Mixtures	MAP + VAD (%)	MVE 5 HMM only + VAD (%)
1.01	128	1	128	4.505	
		8	16	4.502	5.083
	256	1	256	4.127	
		16	16	4.149	3.581
	512	1	512	3.759	
		16	32	3.916	3.182
		32	16	4.036	**3.021**
	1024	1	1024	3.560	
		16	64	4.118	3.163
		32	32	4.301	3.259

Discussion of Evaluation Results

- ~1% absolute EER improvement by MVE training compared to MAP adaptation results for HMM
- ~0.8% absolute EER improvement through modulating the LLR testing results with VAD
- Test setup is consistent with Monitor mode decision period (as opposed to previous NIST utterance based test, which does not apply to Monitor mode)
- Testing decisions made in as little as one second (no system has been reported to work as effectively with this short data)
- Complexity of model increases time to train but does not substantially effect real-time sequential testing

Conclusions

- Contract required work completed
- Unlock software demoed on Samsung Galaxy s4 Smartphone
- Monitor software demoed on Microsoft Surface Pro Tablet
- Submitted final technical report
- Submitted software, user manual, and documentation
- Ready for transition to the Government and real-world applications