EchoPrint:

Two-factor Authentication using Acoustics and Vision on Smartphones

Bing Zhou[1], Jay Lohokare[2], Ruipeng Gao[3], Fan Ye[1]

[1] ECE Department, Stony Brook University

[2] CS Department, Stony Brook University

[3] School of Software Engineering, Beijing Jiaotong University

ACM MobiCom 2018

New Delhi, India

Motivation

PIN

Security issue.

Face Recognition

Image/video spoofing.

Iris Scan

Require special sensors.

Fingerprint Sensor

Take precious space.

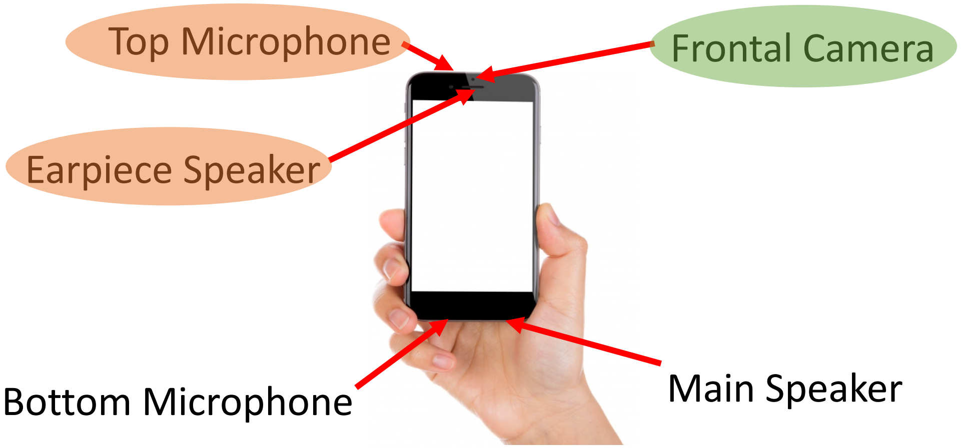
Latest Art

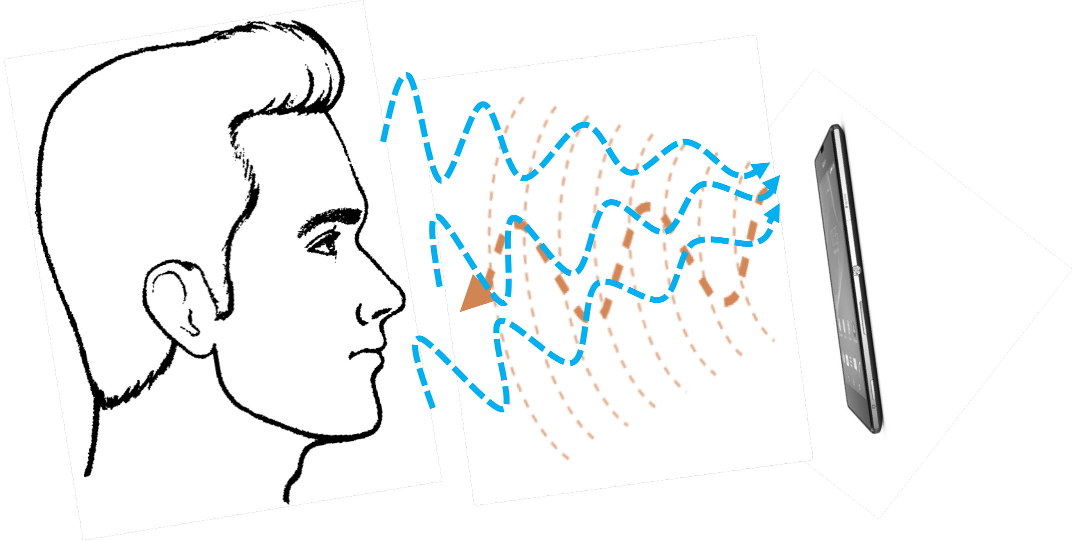
Face ID

High costs, takes precious space.

Is an alternative using existing sensors possible?

Our Approach





Acoustic

- ✓ 3D geometry
- ✓ Sound reflection properties (material)
- ✗ No 2D visual information
- ✗ Highly sensitive to relative pose

+

Vision

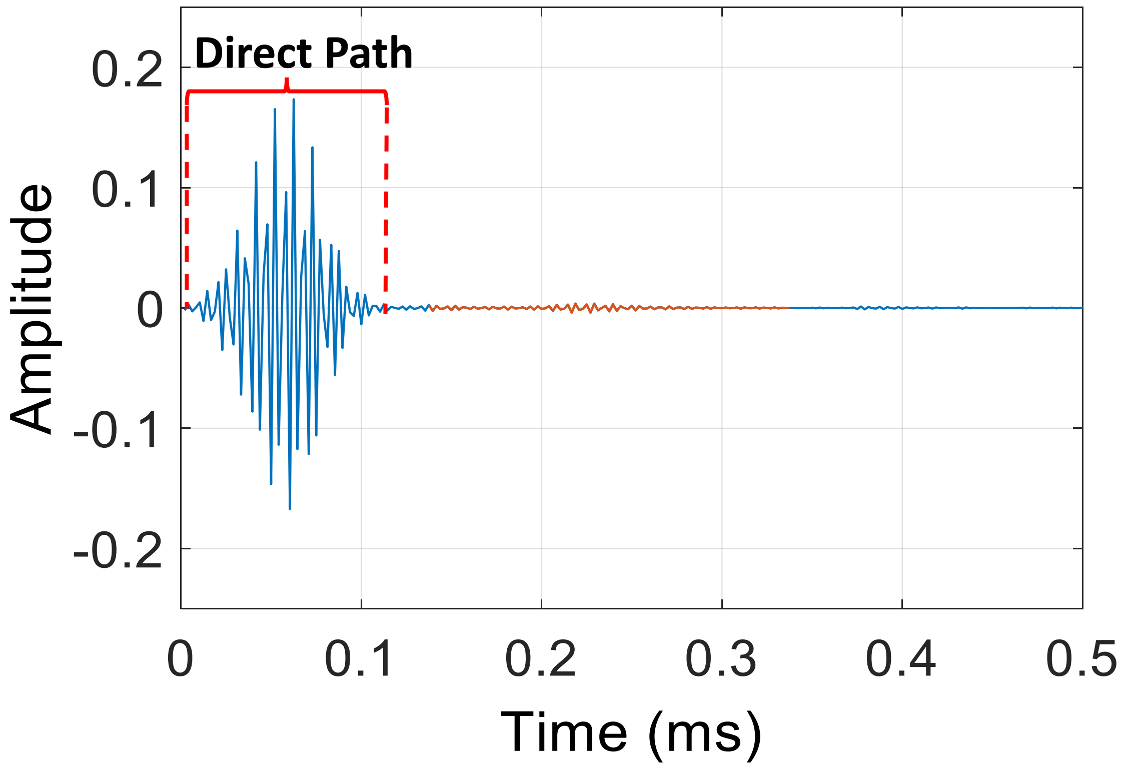
- ✓ Rich 2D appearance information
- ✓ Robust to different angles and distances
- ✗ Can be spoofed by images/videos
- ✗ Subject to lighting conditions

Challenges

- Echo signals are highly noisy and have large variances
  - Hardware limitation of commodity smartphones.
  - Relative pose changes between face and device.
  - ***How do we extract reliable acoustic features despite noise and relative pose changes?***
- Echo signals from face area need to be extracted.
  - Clutters nearby could create even stronger reflections than face.
  - ***How do we segment echo signals from face reliably?***
- Limited training data for user registration
  - Limited data could be collected considering possible relative smartphone poses.
  - ***How do we train a model with limited training samples?***

Acoustic Signal Design

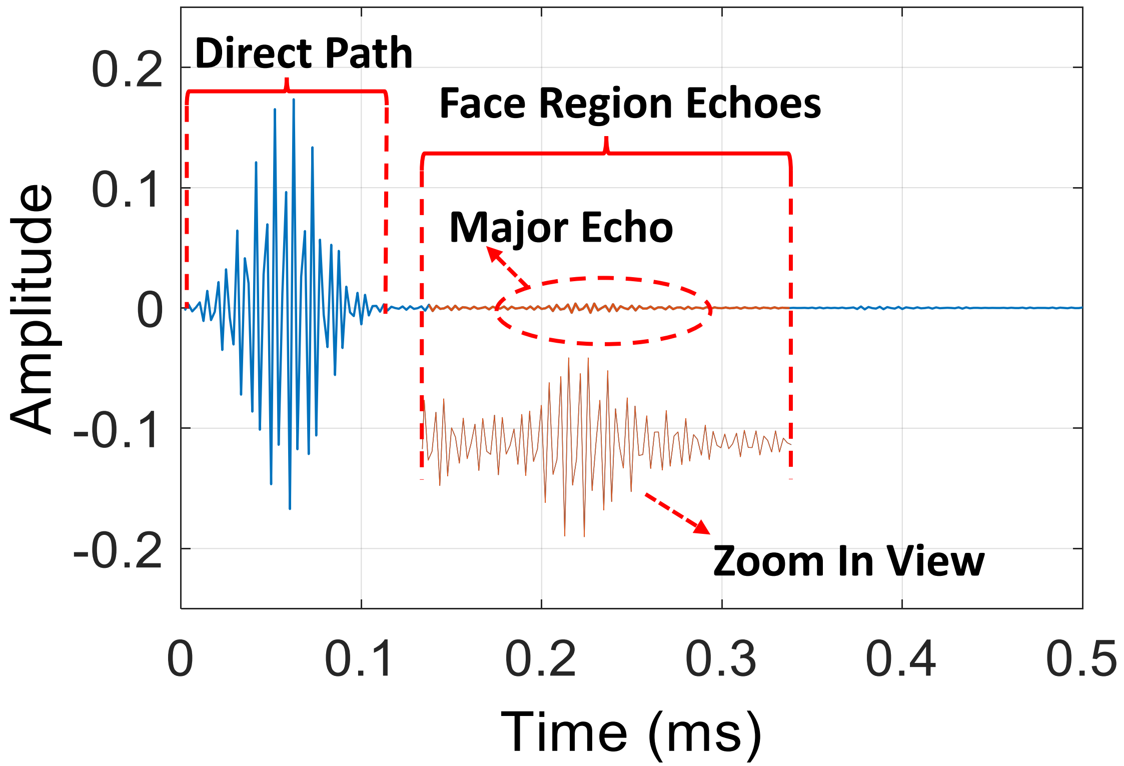
- Pulse signal with a length of 1 ms.
  - Avoid self-interference.
- Linear increasing frequencies from 16 – 22KHz (FMCW).
  - Wide band for higher resolution.
  - Minimize annoyance.
- Reshaped using a Hanning window.
  - Increase peak to side lobe ratio, higher SNR.



Received signal after noise removal.

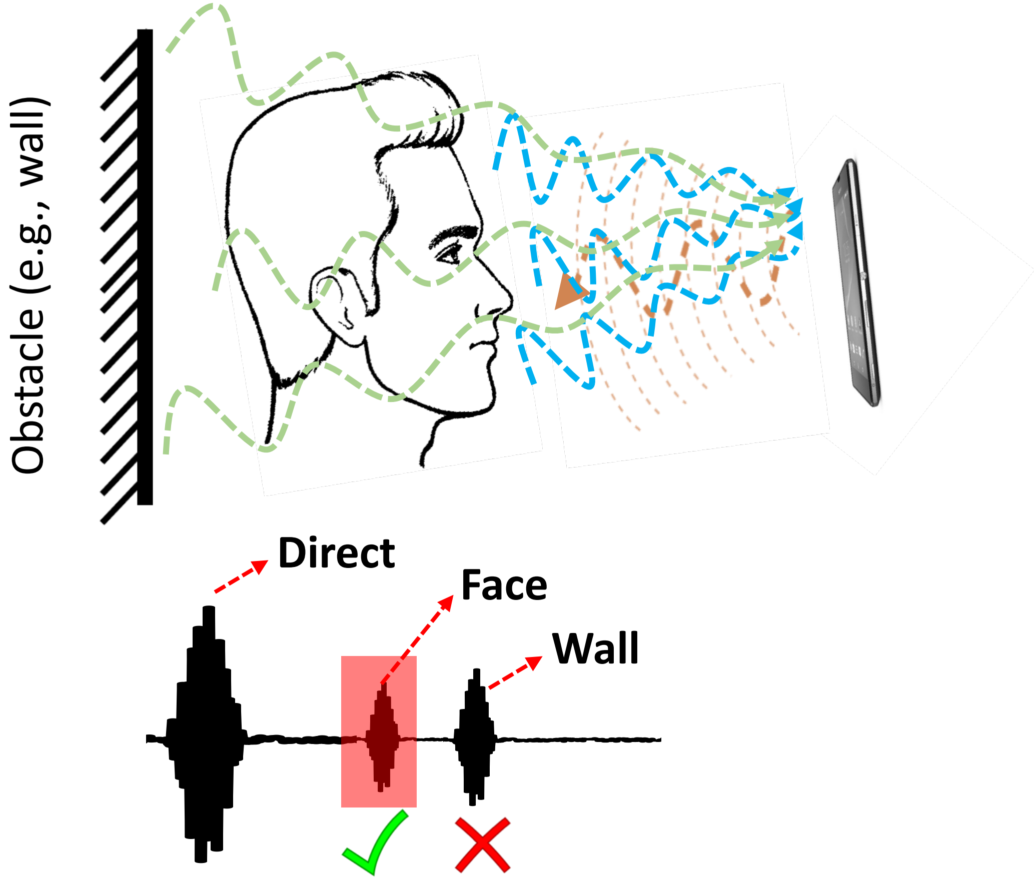
Signal Segmentation

- Background noise removal
  - Butterworth bandpass filter.
- Locate the direct path (Cross-correlation)
  - Template signal calibration
    - Use recorded signal instead of designed signal (hardware imperfection).
- Locate the major echo from face
- Face region echoes
  - Extend 10 sample points before and after major echo (allowing a depth range of ~ 7cm).

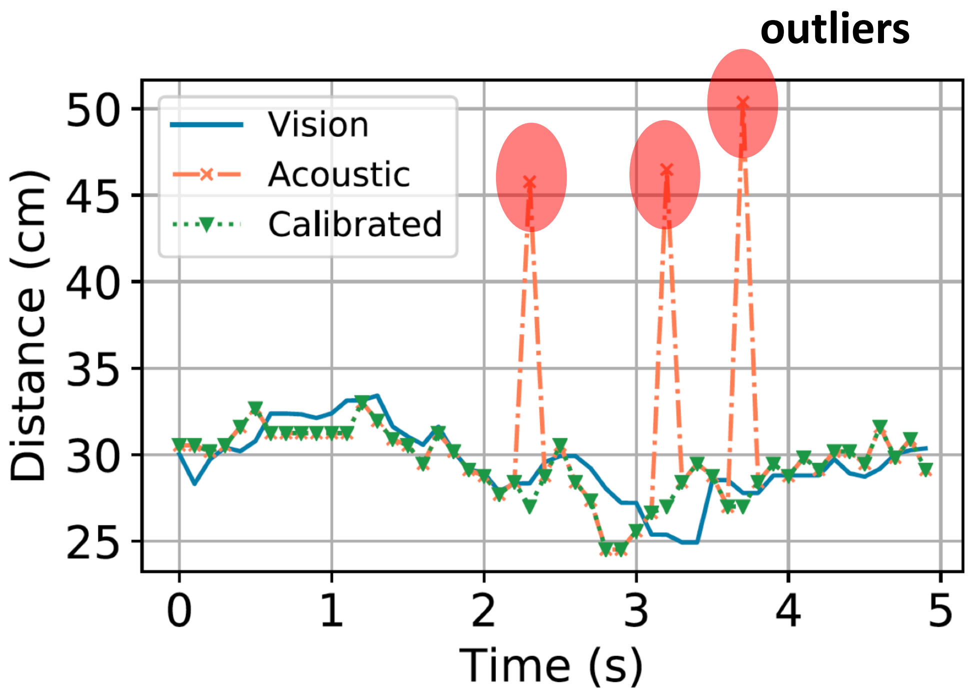


Received signal after noise removal.

Vision-aided Major Echo Locating



How do we tell which one is from face?

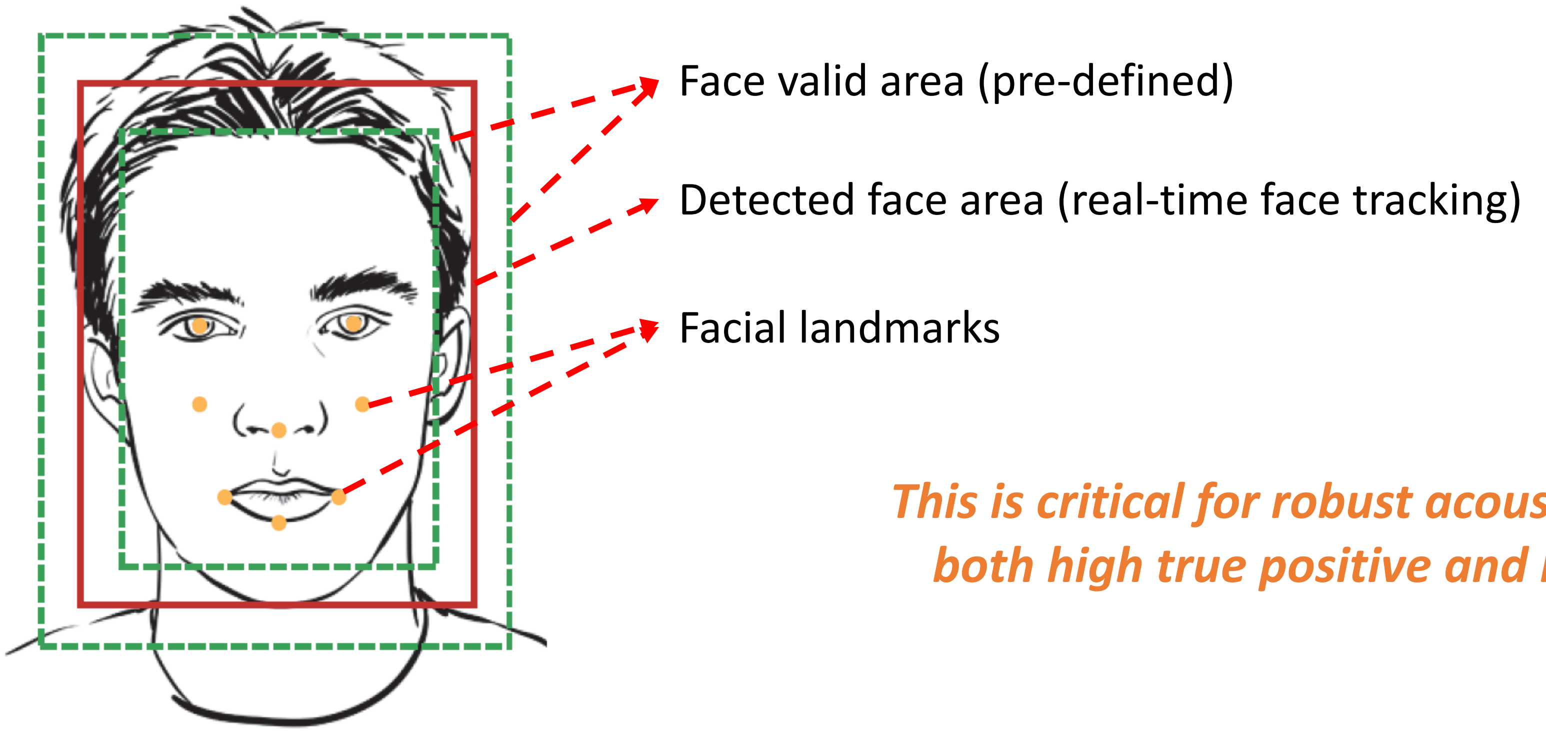


Vision: **rough but robust** distance estimates from landmarks.
Acoustic: **accurate but outliers** may exist.

Leverage vision measurements to narrow down the "search" region of acoustic echoes.

Face Alignment

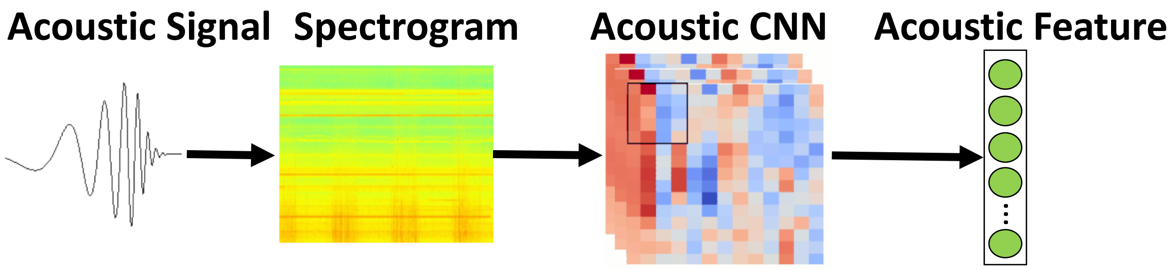
- Real-time face tracking and facial landmark detection on mobile
  - Face tracking is used for face alignment, thus confining the relative pose.
  - Landmarks are used for distance estimation, helping major echo locating.



This is critical for robust acoustic sensing, enabling both high true positive and low false negative.

Acoustic Representation Learning

- CNN model for feature extraction
  - Input: spectrogram after FMCW mixing.
  - Output: 128 dimensional feature vector.
  - 710593 parameters.
- Trained on a data set of 91708 valid samples from 50 subjects.
- Last layer was removed to be used as feature extractor.



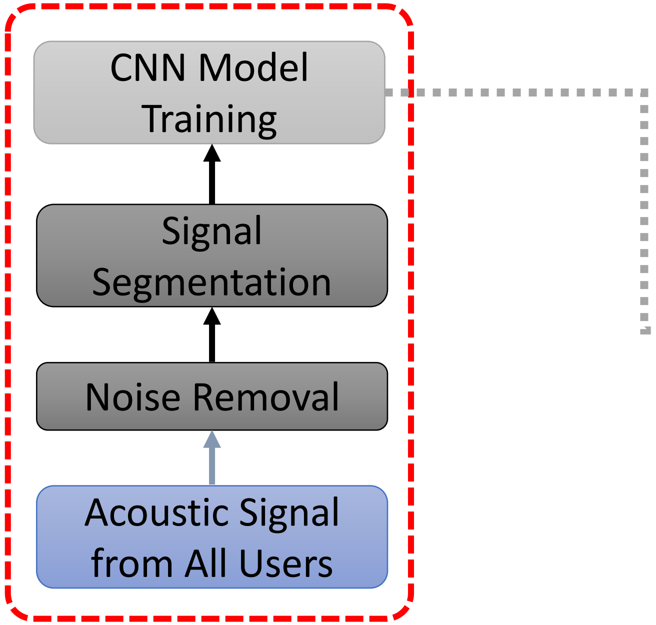
Layer	Layer Type	Output Shape	# Param
1	Conv2D + ReLU	(33,61,32)	320
2	Conv2D + ReLU	(31,59,32)	9248
3	Max Pooling	(15,29,32)	
4	Dropout	(15,29,32)	
5	Batch Normalization	(15,29,32)	128
6	Conv2D + ReLU	(15,29,64)	18496
7	Conv2D + ReLU	(13,27,64)	36928
8	Max Pooling	(6,13,64)	
9	Dropout	(6,13,64)	
10	Batch Normalization	(6,13,64)	256
11	Flatten	(4992)	
12	Dense + ReLU	(128)	639104
13	Batch Normalization	(128)	512
14	Dense + Softmax	(50)	5547

CNN architecture.

Authentication Model

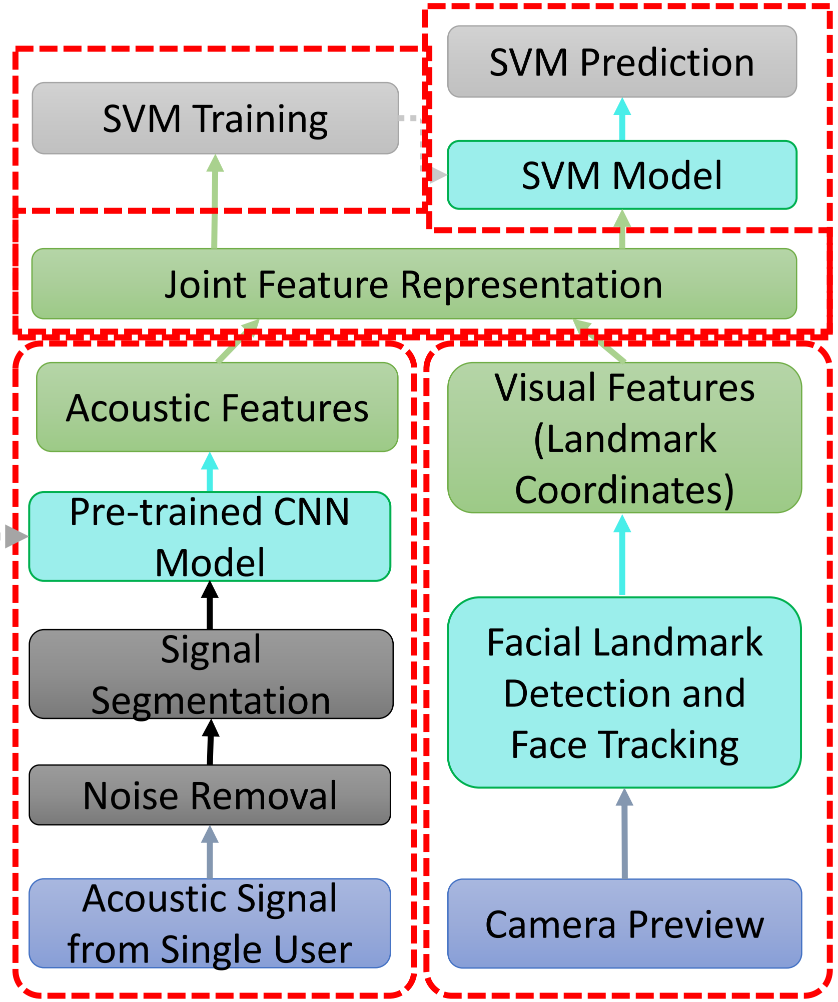
Acoustic Representation Learning

(CNN ***one-time off-line*** training **on PC**)



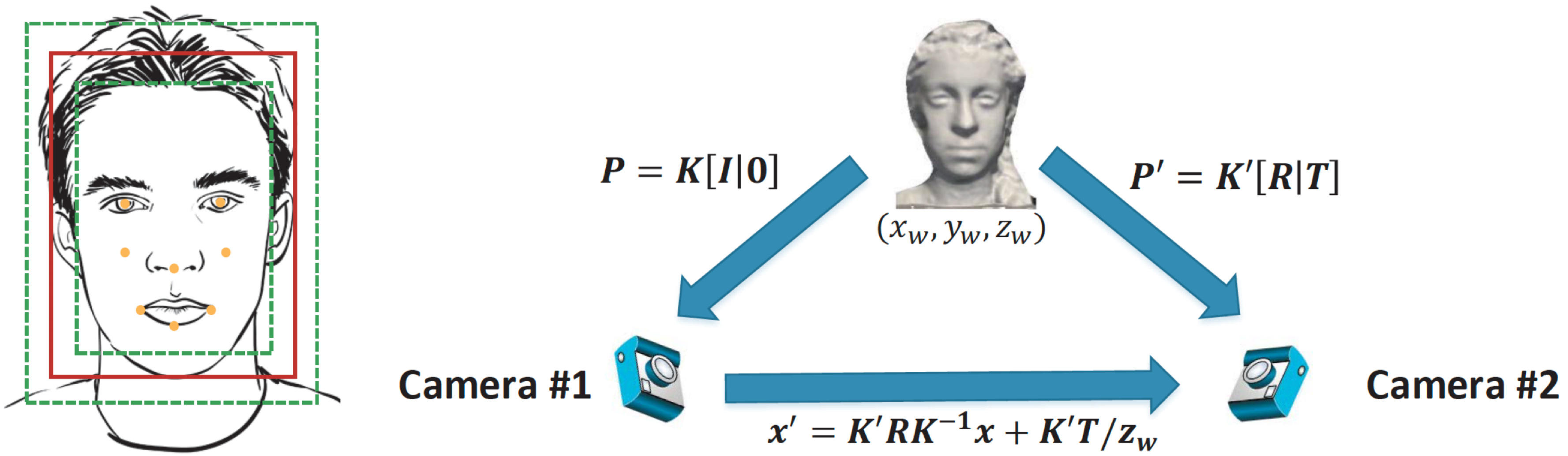
Two-factor Authentication

(SVM training and ***real-time*** prediction **on smartphones**)



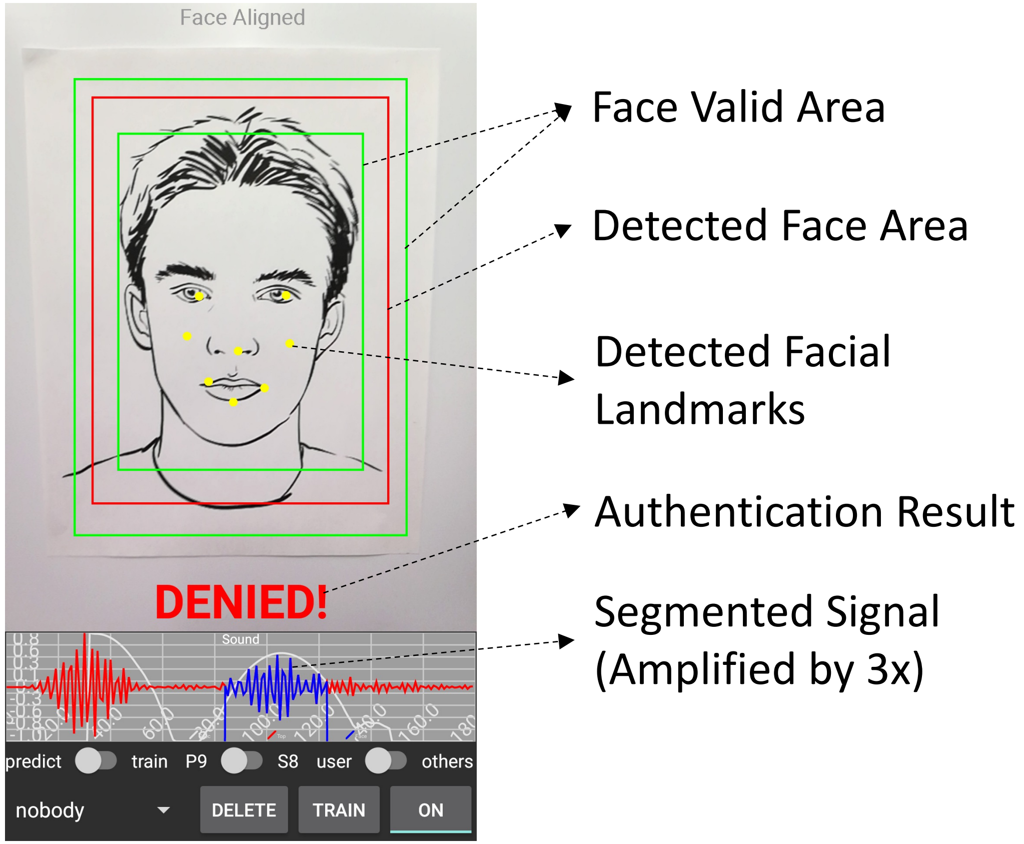
Data Augmentation

- **Populate the training data** by generating **"synthesized" training samples** based on *facial landmark transformation* and *acoustic signal prediction*.
  - Step 1: Compute the landmark's world coordinates.
  - Step 2: Transform the landmark onto new images, assuming the camera is at a different pose.
  - Step 3: Adjust acoustic signal according to the sound propagation law.
  - Step 4: Generated landmarks and acoustic signal form a "synthesized" training sample.



Implementation

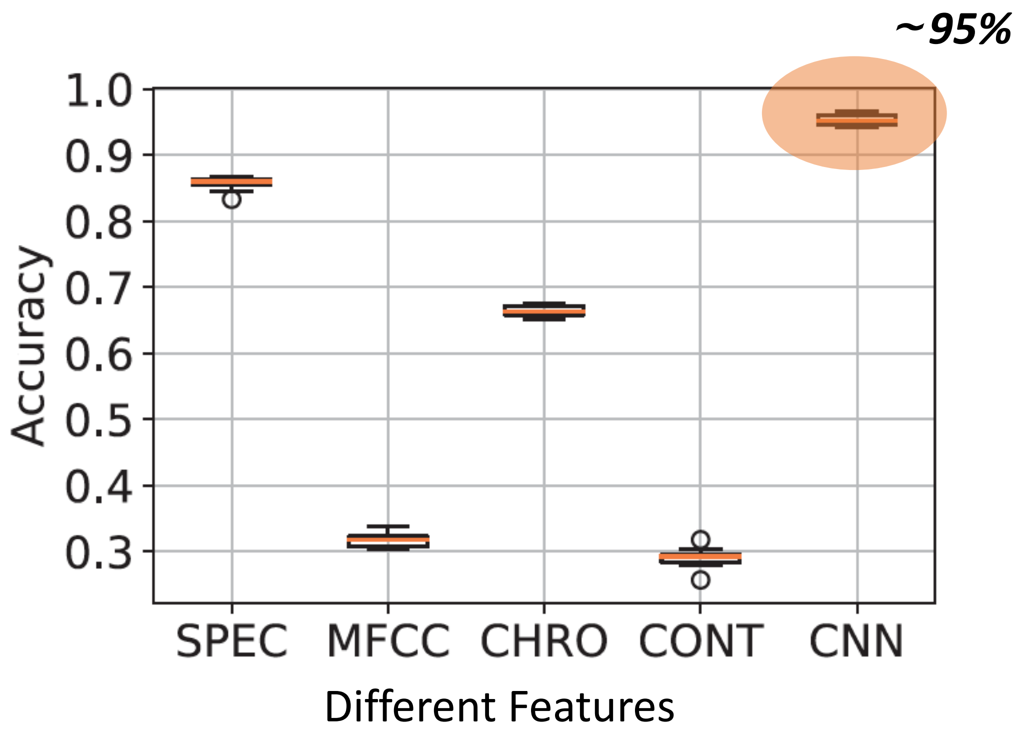
- Android prototype
  - Face tracking and landmark detection.
    - Google mobile vision API
  - Acoustic sensing pipeline.
    - Android SDK
  - On-device machine learning pipeline.
    - LibSVM, TensorFlow
- Offline CNN training
  - CNN trained offline on a PC with GTX 1080 Ti GPU.
  - Pre-trained CNN model was frozen and deployed on mobile device.

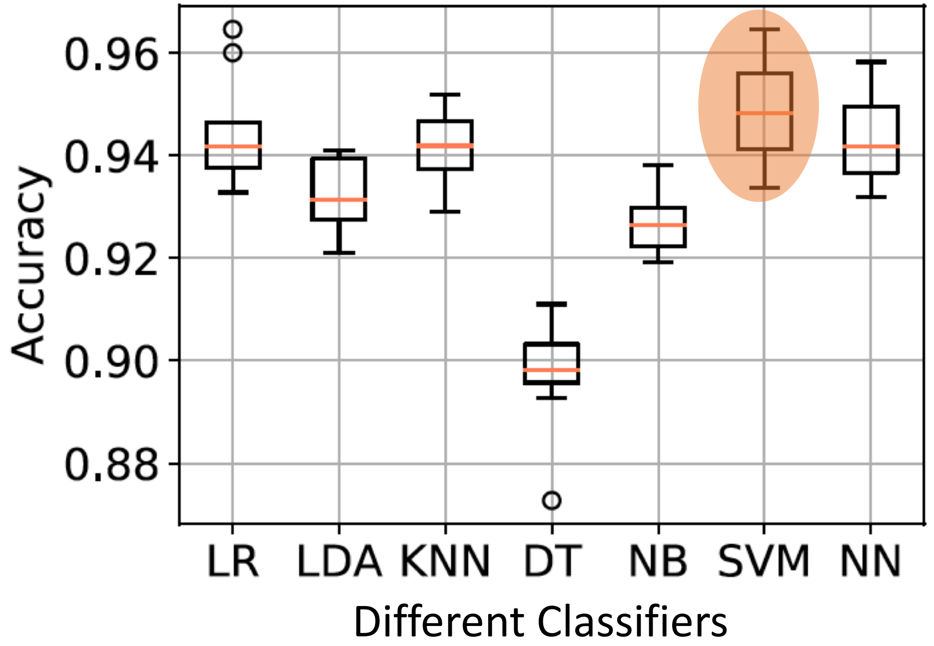


Evaluations --- Data Collection

- Data source
  - ***45 participants*** of different ages, genders, and skin colors
  - ***5 non-human classes***:
    - Photos, monitors, tablets, marble sculptures, etc.…
- Data collection rule
  - Move the phone slowly to cover different poses.
  - Multiple uncontrolled environments (quiet lab, noisy classroom, outdoor).
  - Different lighting conditions.
  - Multiple sessions at different times and locations.
- Data amount
  - 120 Seconds, 7-8 MB data, ~2000 samples for each subject.
  - 91708 valid samples from 50 classes, 70% for training, 15% each for model validation and testing.
  - Additionally, 12 more volunteers join as ***NEW users*** for evaluation.

Evaluations --- CNN Feature Extractor





SPEC: Spectrogram
MFCC: Mel-Frequency Cepstral Coefficients
CHRO: Chromagram
CONT: Spectral Contrast

LR: Linear Regression
LDA: Linear Discriminant Analysis
KNN: K-nearest Neighbor
DT: Decision Tree
NB: Naïve Bayesian
SVM: Support Vector Machine
NN: Neural Network

Evaluations --- Performance on New Users

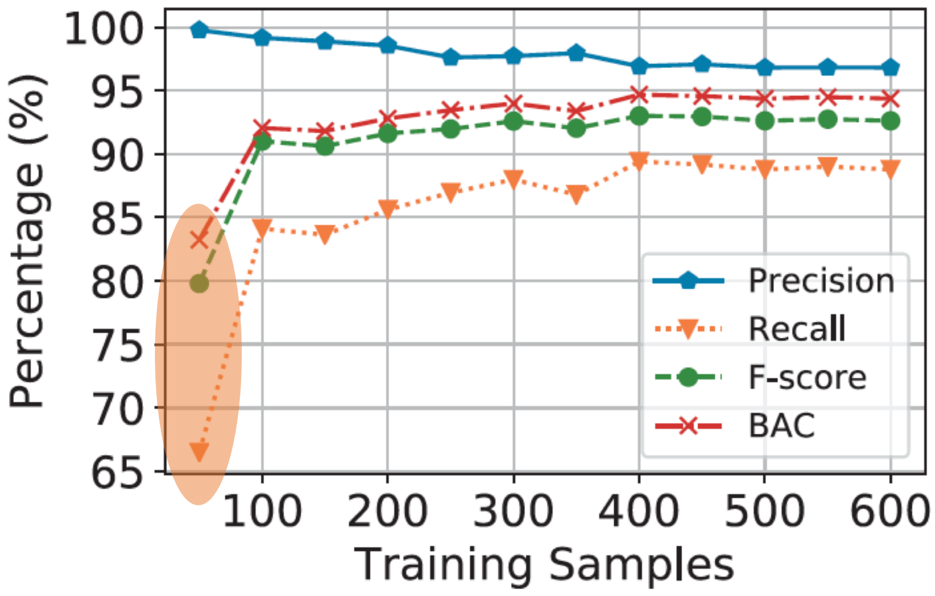
- 12 volunteers (data not used in CNN training)
  - ~2 minutes data, half for training, and half for testing.
- Metrics
  - Precision: the higher, the less false positive, the more secure.
  - Recall: the higher, the less false negative, more user friendly.

$$P = \frac{TP}{TP+FP}$$

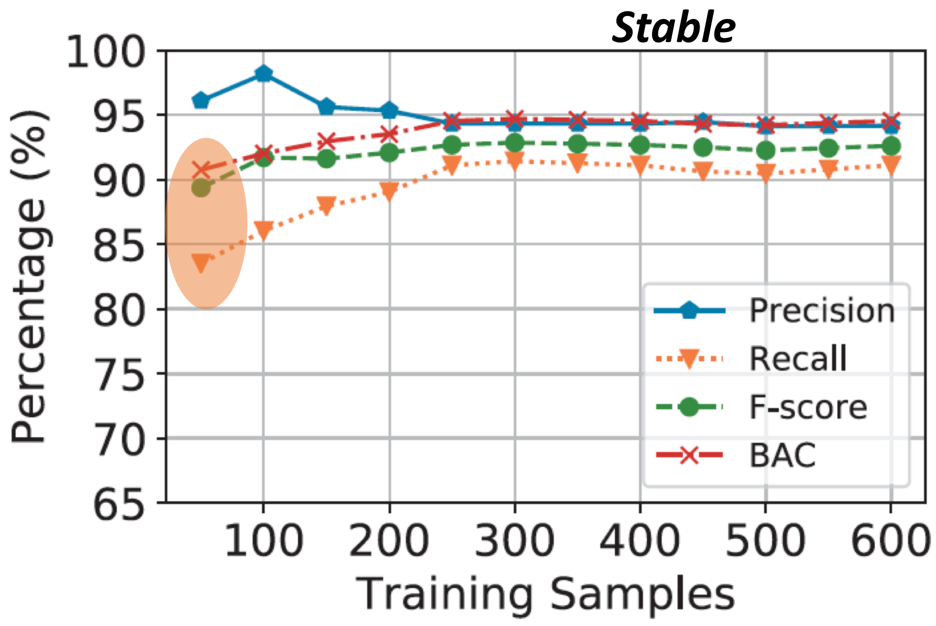
$$R = \frac{TP}{TP+FN}$$

	Mean	Median	Standard Deviation
Precision (%)	98.05	99.21	2.78
Recall (%)	89.36	89.31	1.62
F-Score (%)	93.50	94.33	1.68
BAC (%)	93.75	94.52	0.85

Evaluations --- Data Augmentation



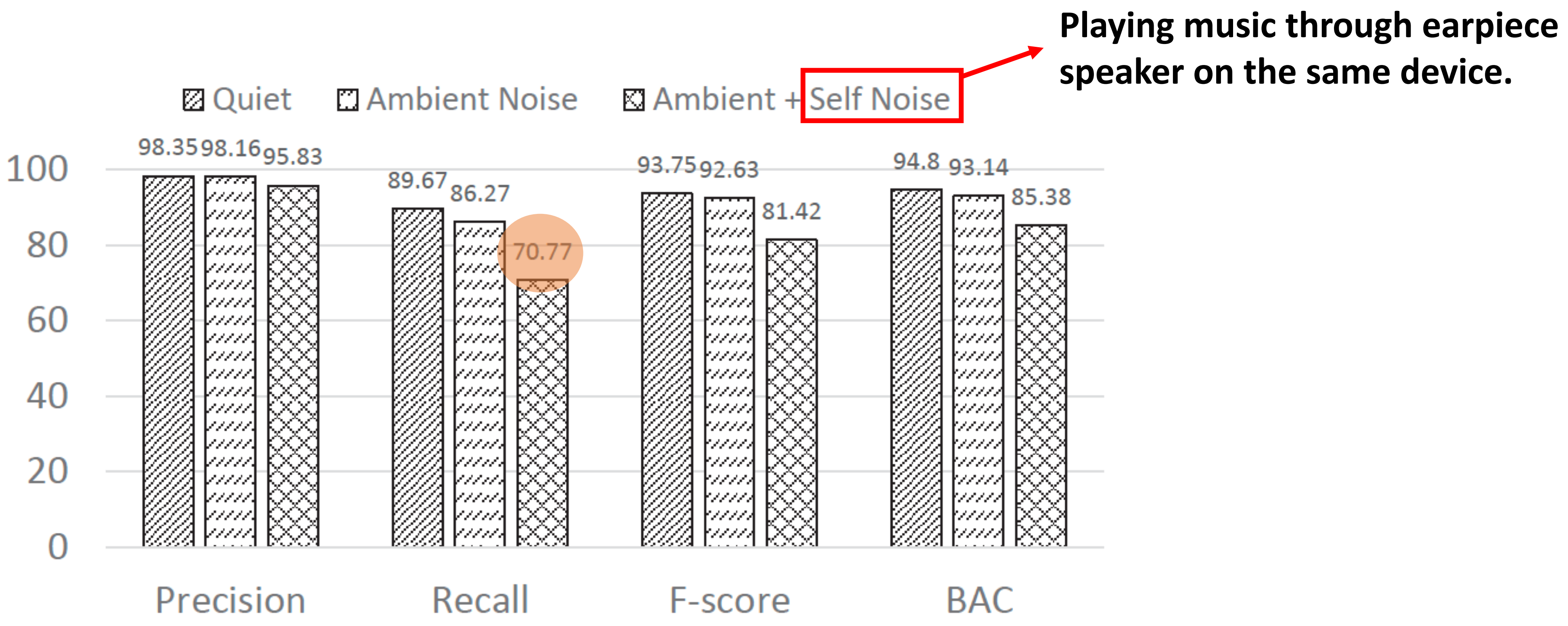
Without data augmentation.



With data augmentation.

Data augmentation improves recall significantly when the training samples are very limited.

Evaluations --- Background Noise



Performance under difference noises.

Background noise does not have obvious impact on performance.

Evaluations --- Image Spoofing

- Spoofing attacks
  - Color photos of 5 volunteers in 10 different sizes on paper.
  - Display the photos on desktop monitors while zooming in/out gradually.
  - Various distance between 20 – 50 cm.
- They easily pass pure vision face recognition based system [1], but all failed our two-factor authentication.

[1] Amos, B., Ludwiczuk, B. and Satyanarayanan, M., 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*.

Evaluations --- Resource Consumption

- Memory & CPU consumption & response delay

Device	Memory (MB)	CPU (ms)	Delay (ms)
Samsung S7	22.0 / 50.0	6.42 / 31.59	44.87 / 91
Samsung S8	20.0 / 45.0	5.14 / 29.04	15.33 / 35
Huawei P9	24.0 / 53.0	7.18 / 23.87	32.68 / 86

Mean / max resource consumption.

Small amount of memory

Real-time recognition

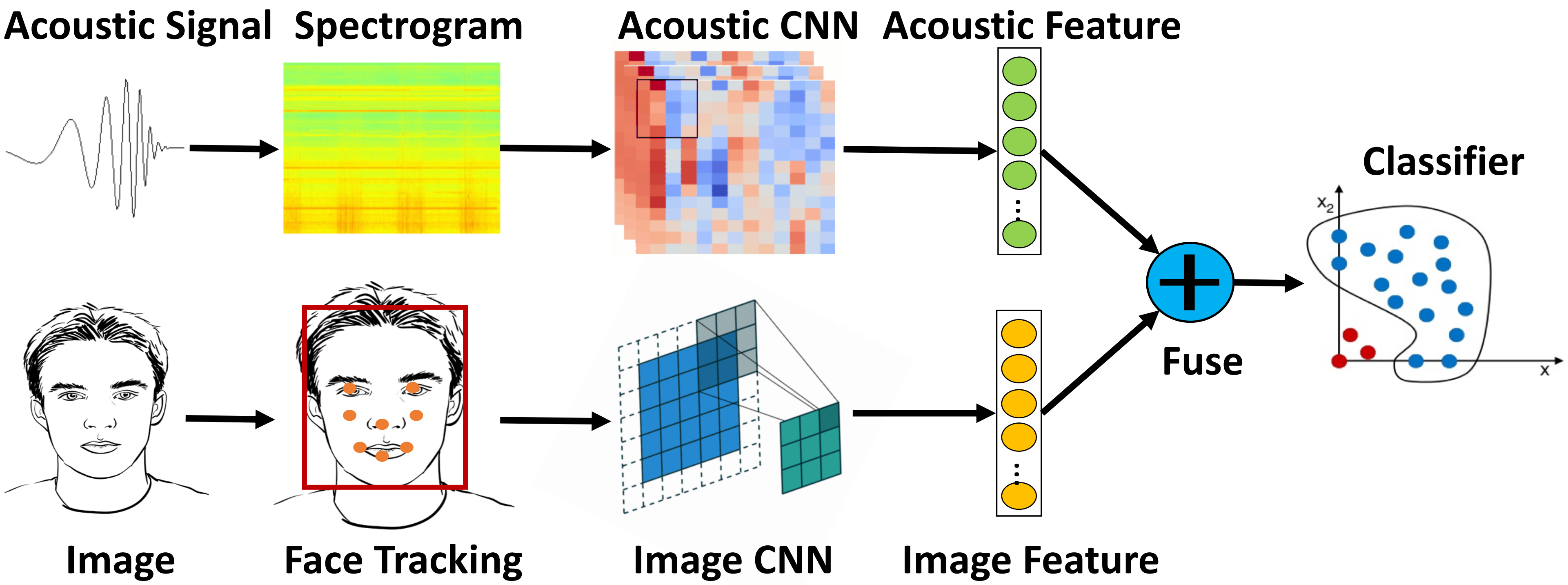
Unobvious delay

Limitations

- Requirement of face alignment
  - Inconvenient for daily use.
- Limitations from vision
  - Face tracking is not stable under poor lighting.
- User appearance changes
  - Online model updating mechanism is needed.
- Continuous authentication usability
  - Limited usability due to face alignment.

Working Progress

- Leveraging sophisticated visual features
  - e.g. OpenFace [1]
  - Less constraints on face alignment, better usability, higher accuracy.



[1] Amos, B., Ludwiczuk, B. and Satyanarayanan, M., 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*.

Future work

- Enhancing CNN acoustic feature extractor
  - More data from more users with larger variety.
  - More sophisticated neural network design.
- Integration with existing solutions
  - Integrated with existing commercial authentication solutions.
- Large scale experiment
  - Large scale experiment (e.g., thousands or more) is needed for a mature solution.

Thank You.

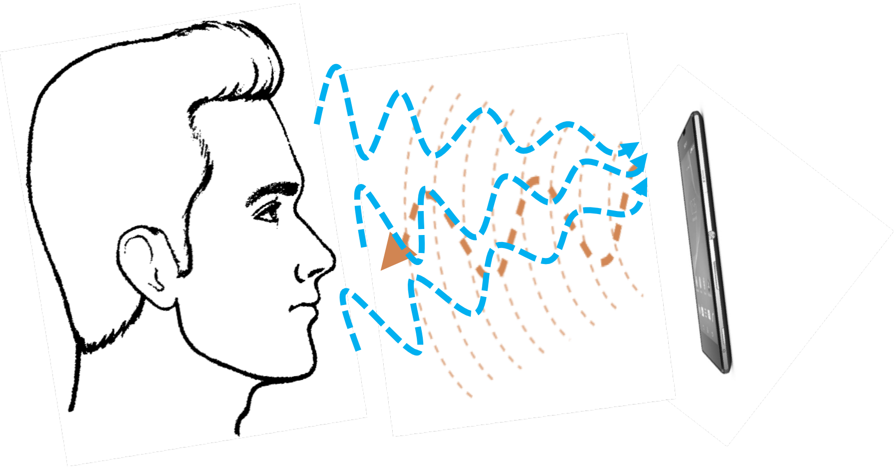
Backup Slides

Design Considerations

- Universal
  - Use existing hardware on most smartphones
  - Use a biometric that is pervasive to every human being.
- Unique
  - Distinctive biometric (2D visual based systems can be spoofed easily).
- Persistent
  - Biometric must not change much over time (heart beat, breathing, gait are highly affected by physical conditions).
- Difficult to circumvent
  - Circumventing require duplicating both 3D facial geometries and acoustic reflection properties close enough to human face.

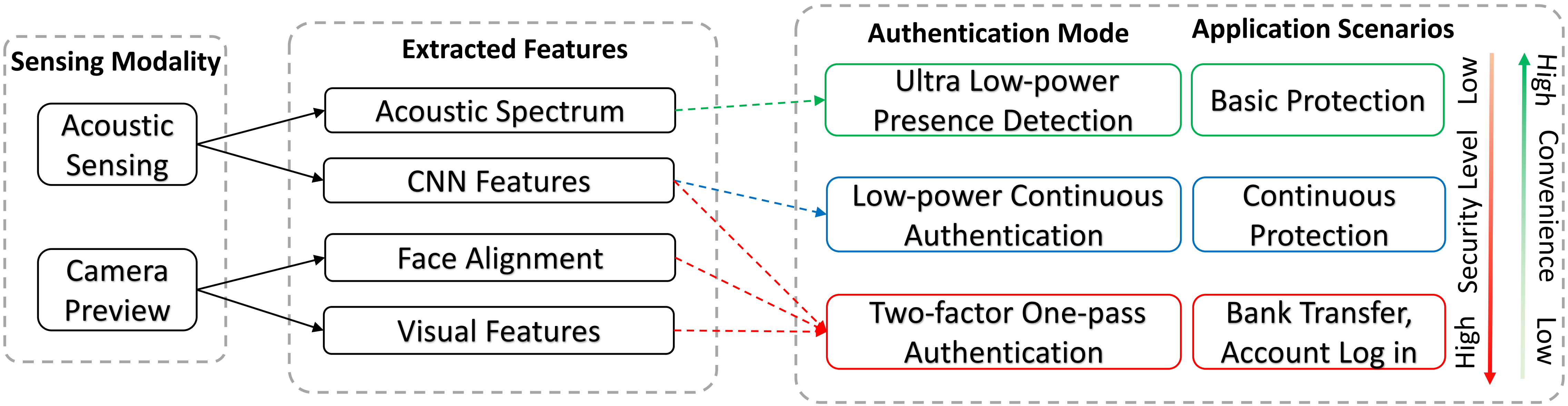
Our Approach

- Acoustic signal
  - Low propagation speed
    - High ranging accuracy
  - Light computation
    - Orders of magnitude less compared to vision method
  - Existing hardware
    - Almost all smart devices have speakers and microphones



Authentication Modes

- Two-factor one-pass authentication
- Low-power continuous authentication
- Ultra low-power presence detection



Evaluations --- Authentication Accuracy

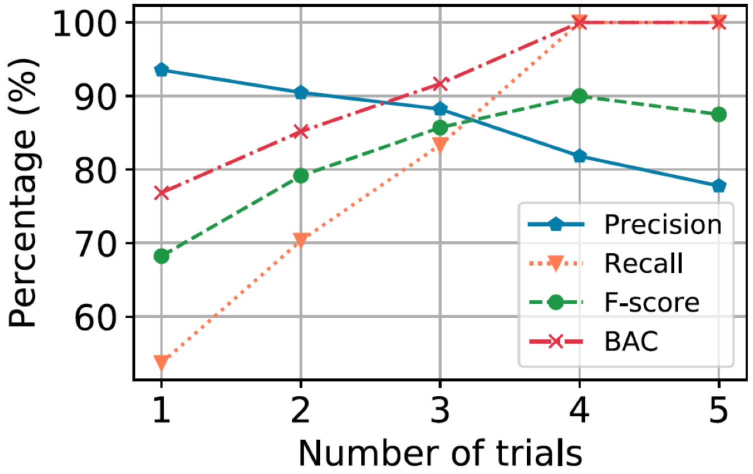
- Precision, Recall and BAC

Table 2: Mean/median accuracy with vision, acoustic and joint features.

	Vision	Acoustic	Joint
Precision (%)	72.53 / 80.32	86.06 / 99.41	88.19 / 99.75
Recall (%)	64.05 / 64.04	89.82 / 89.84	84.08 / 90.10
F-score (%)	65.17 / 69.19	85.39 / 94.31	83.74 / 93.23
BAC (%)	81.78 / 81.83	94.79 / 94.88	91.92 / 95.04

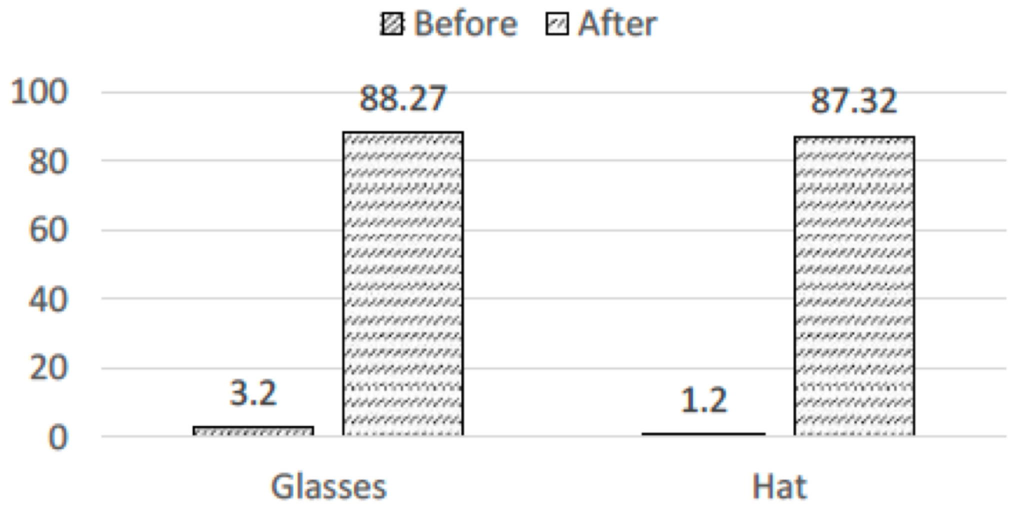
Evaluations --- Continuous Modes

- Continuous authentication using acoustic only
  - The volunteer tries to keep the face aligned while camera is disabled.
  - One verdict from multiple trials.
- Still have usability issue
  - Users are unlikely to keep face aligned while using the device.



Continuous authentication performance with different number of trials.

Evaluations --- User Appearance Changes



Average recall of 5 users before/after model updating with new training data.

Evaluations --- Resource Consumption

Power consumption

Device	ULP (mW)	LP (mW)	Two-factor (mW)	Vision (mW)
S7	305	1560	2485	1815
S8	215	1500	2255	1655
P9	265	1510	2375	1725