Home Assistant Sonar

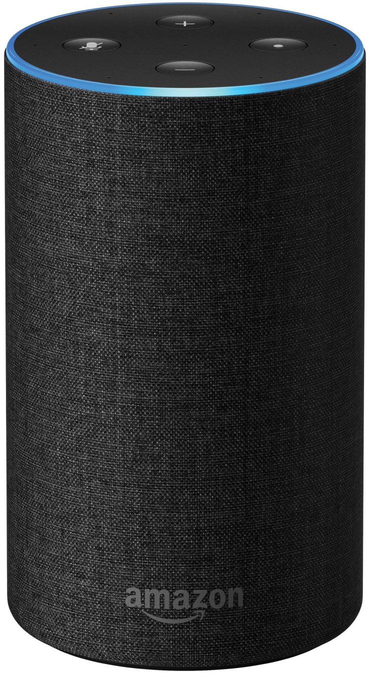
Dennis Shim, Seraphine Goh

EE209AS
June 14, 2019

Introduction

OK Google

Alexa



Adversarial attacks on smart speakers abuse programmed “hot word” and active microphone

Problem Statement (1)

- Case 1: Binary
  - Identify whether or not human is present in room

Problem Statement (2)

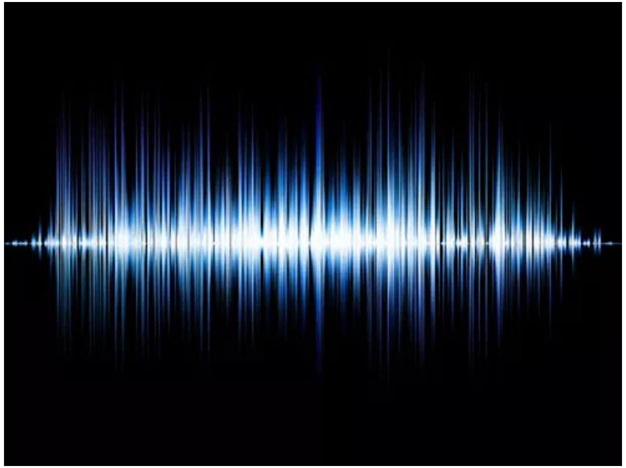
- Case 2: Multi-Class
  - Identify human’s location in room, with respect to smart speaker
  - In “front”, “behind”, to the “right”, or to the “left”

Prior Work

- EchoSafe
  - Detect user presence
  - Using 1 kHz sonar
  - Binary Random Forest (RF) classifier
  - Feature selection using Relief-F algorithm
  - Achieved 93.13% accuracy

- Automatic Speaker Verification (ASV)
  - Voice fingerprinting

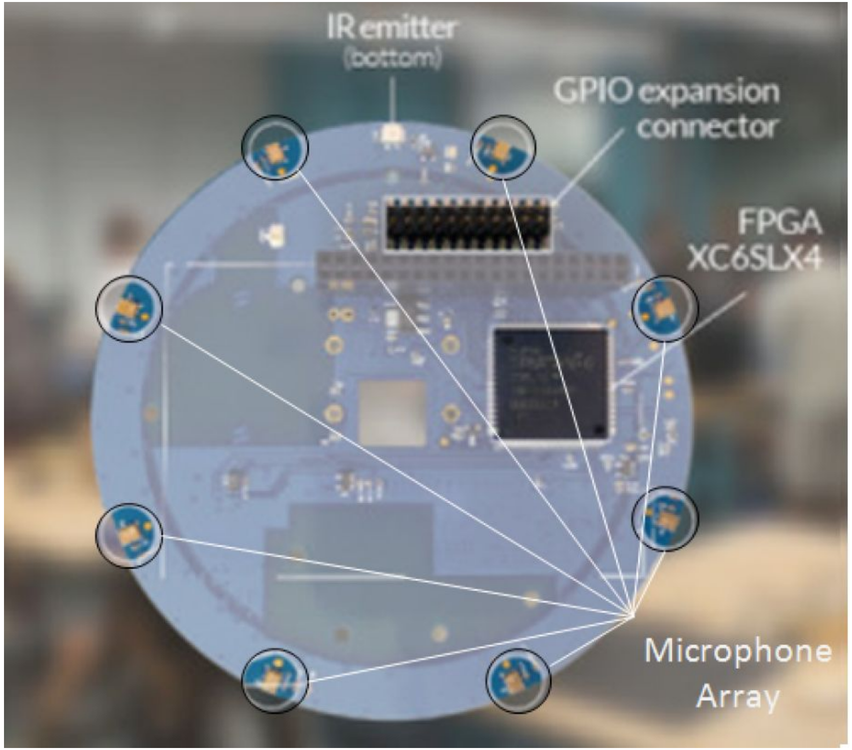
- Audio beamforming
  - Get angle of incoming sound with multiple microphones



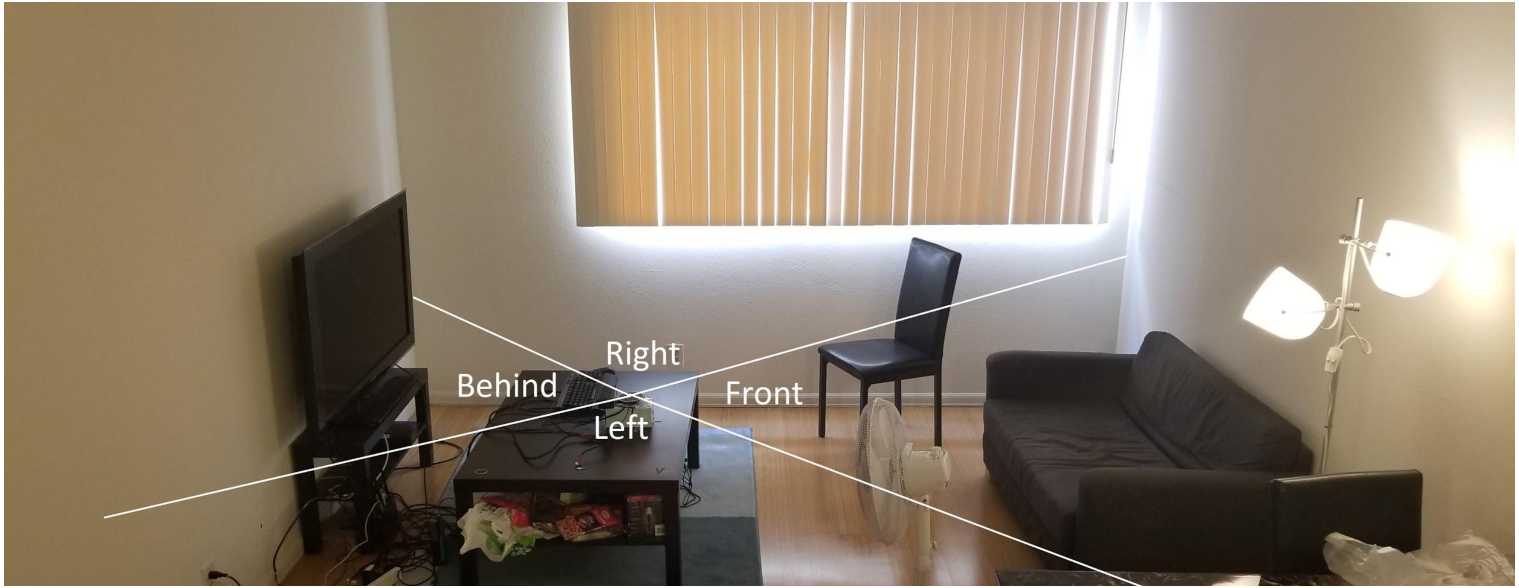
Technical Approach: Data Collection (1)

- Matrix Creator
  - Open source home assistant
  - Eight-microphone array
  - 20 Hz to 20 kHz, beamforming
  - FPGA

- Raspberry Pi 3
  - Connects to Matrix Creator via GPIO
  - Used for data collection



Technical Approach: Data Collection (2)



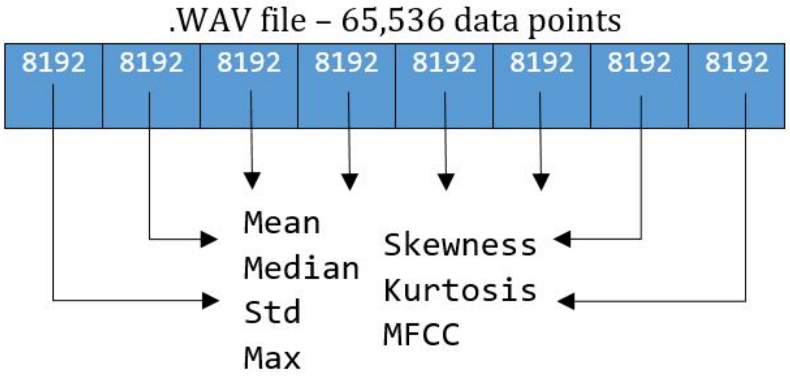
- Data was collected from a living room with the Matrix Creator set up in the center
- Data was labelled according to the picture

Technical Approach: Data Collection (3)

Frequency of Tone	Subject Position	# of trials	Label (Binary Label)
1 kHz	No subject	455	0 (0)
	Front (couch-side)	400	1 (1)
	Behind (TV-side)	100	2 (1)
	Right (window-side)	200	3 (1)
	Left (front door-side)	200	4 (1)
20 kHz	No subject	210	0
	Front (couch-side)	200	1
100 Hz	No subject	210	0
	Front (couch-side)	210	1
1 kHz	Moving counterclockwise	100	1
	Moving clockwise	100	2

Technical Approach: Feature Extraction

- Each trial generates nine 4-second .WAV files
  - One for each mic (8) and one beamformed
  - 65,536 data points per file
- Each file was split into 8 non-overlapping windows
- For each window, features were extracted:
  - Mean, median, standard deviation, max, skewness, kurtosis, 70 MFCCs
  - MFCCs capture features of the frequency spectrum



Analysis and Results: Feature Selection

Setup:

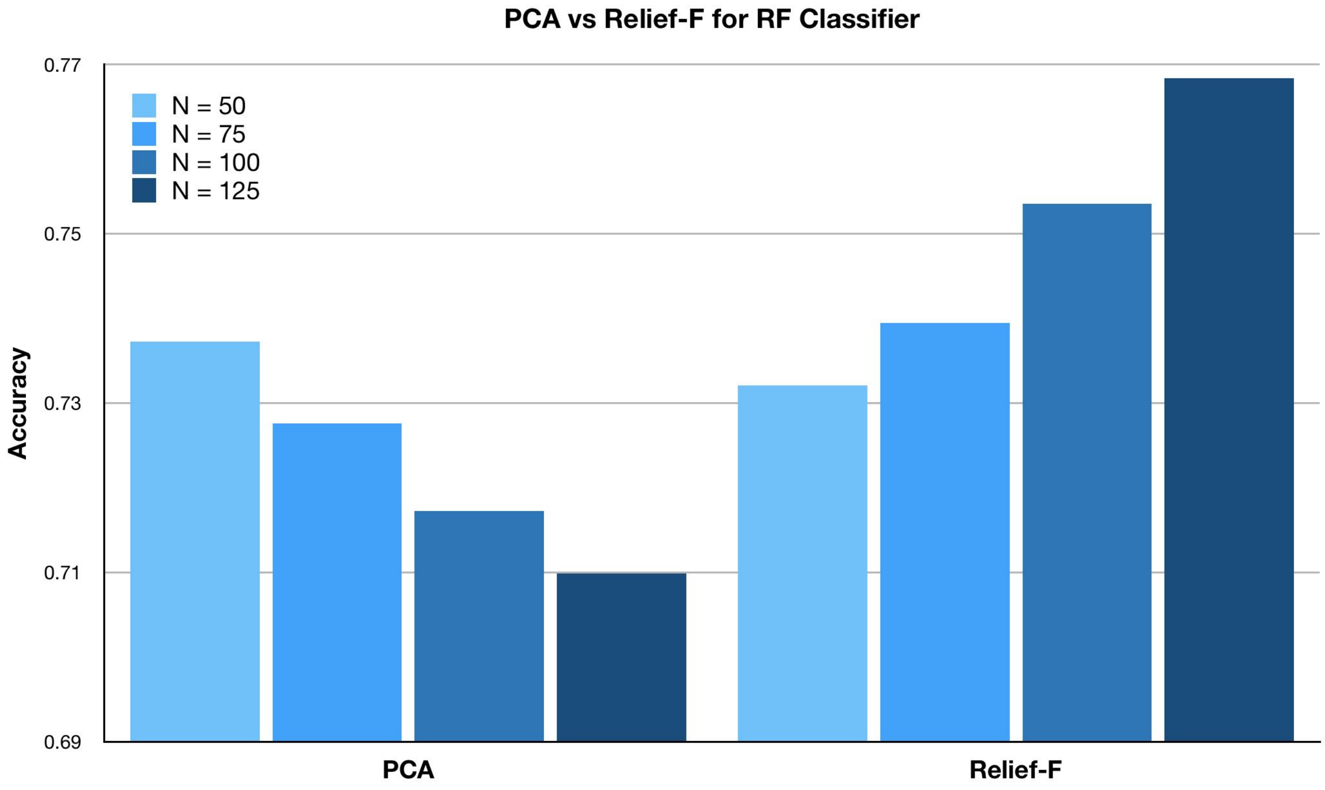
- PCA vs Relief-F
- Binary Random Forest (RF) classifier
- Vary N features

Results:

- Relief-F obtained overall higher accuracies than PCA

Conclusion:

- Use Relief-F for feature selection



Analysis and Results: ML Classification (1)

Setup:

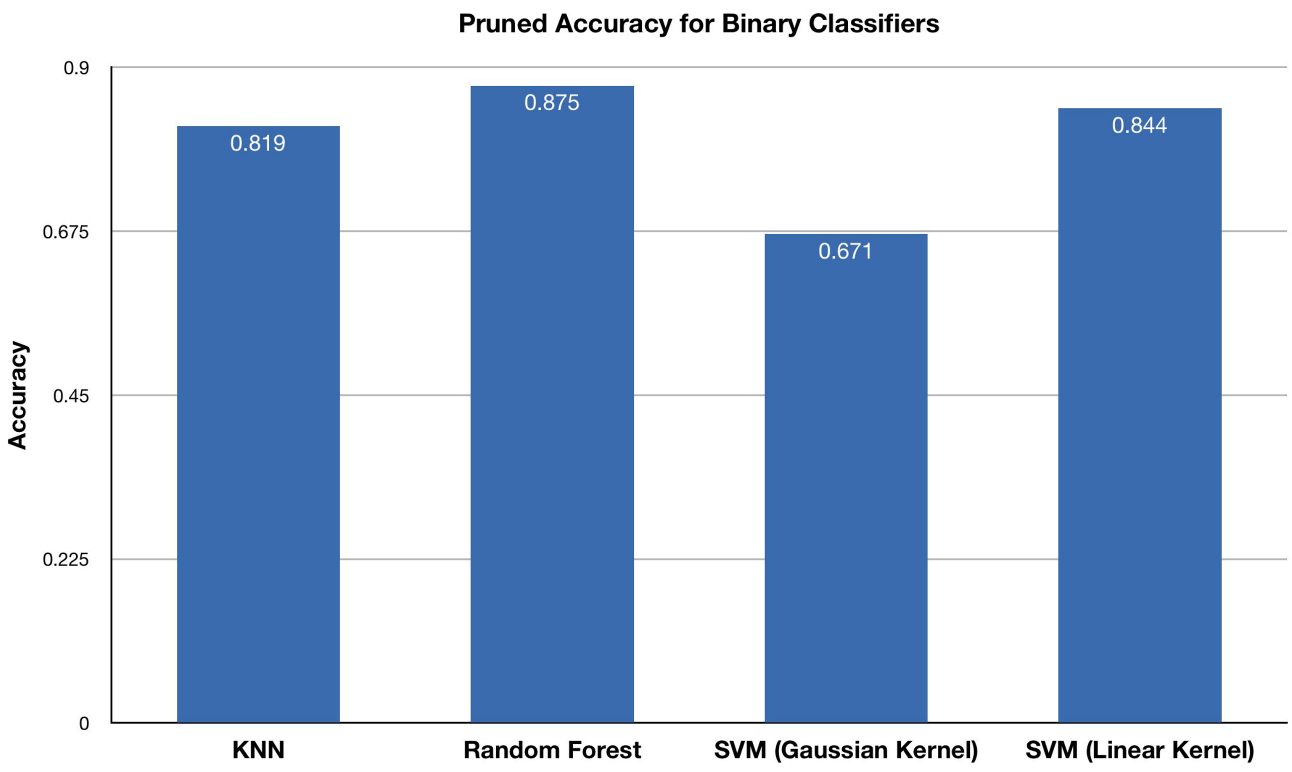
- Binary classifiers
- N = 125 features
- Pruned w/ Relief-F

Results:

- Cross-validation accuracies for KNN, RF, KNN, SVM (Gaussian/Linear)

Conclusion:

- Top 3: RF, KNN, SVM (Linear)



Analysis and Results: ML Classification (2)

Setup:

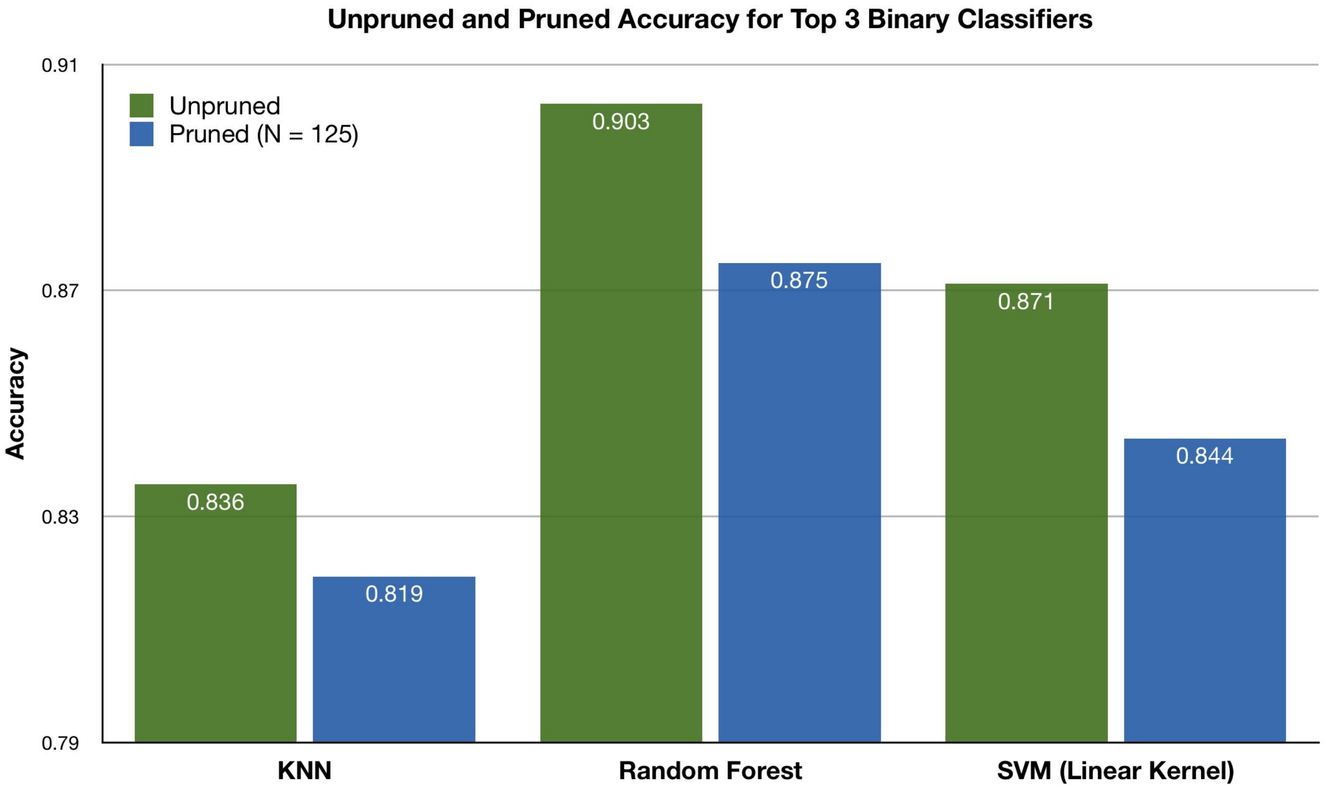
- Top 3 binary
- Unpruned (N = 5462 features)
- Pruned (N = 125)

Results:

- Compare unpruned & pruned accuracy

Conclusion:

- Use binary RF for best cross-validation accuracy

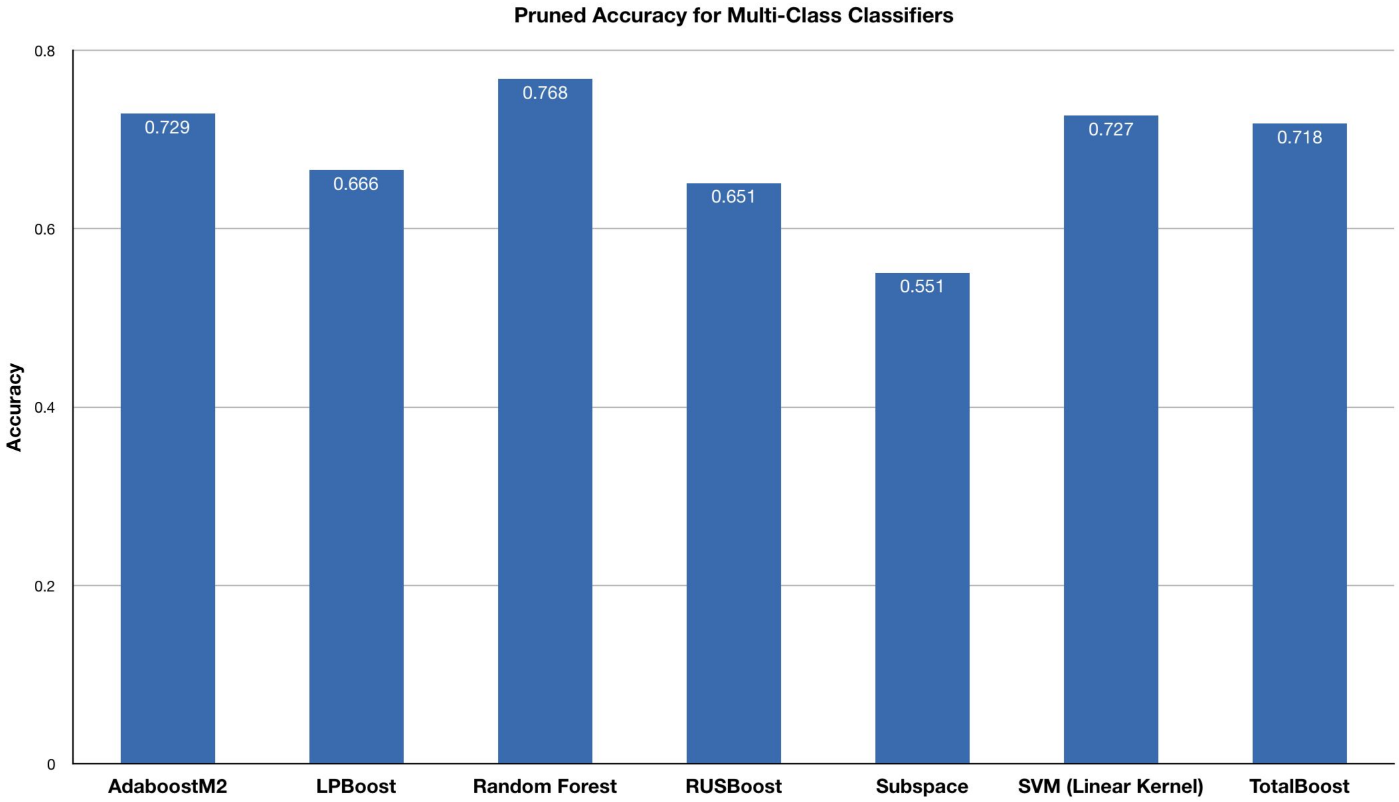


Analysis and Results: ML Classification (3)

Setup:
- N = 125 features
- Pruned w/ Relief-F

Results:
- Cross-validation accuracies

Conclusion:
- Top 3: RF, AdaboostM2, SVM (Linear)



Analysis and Results: ML Classification (4)

Setup:

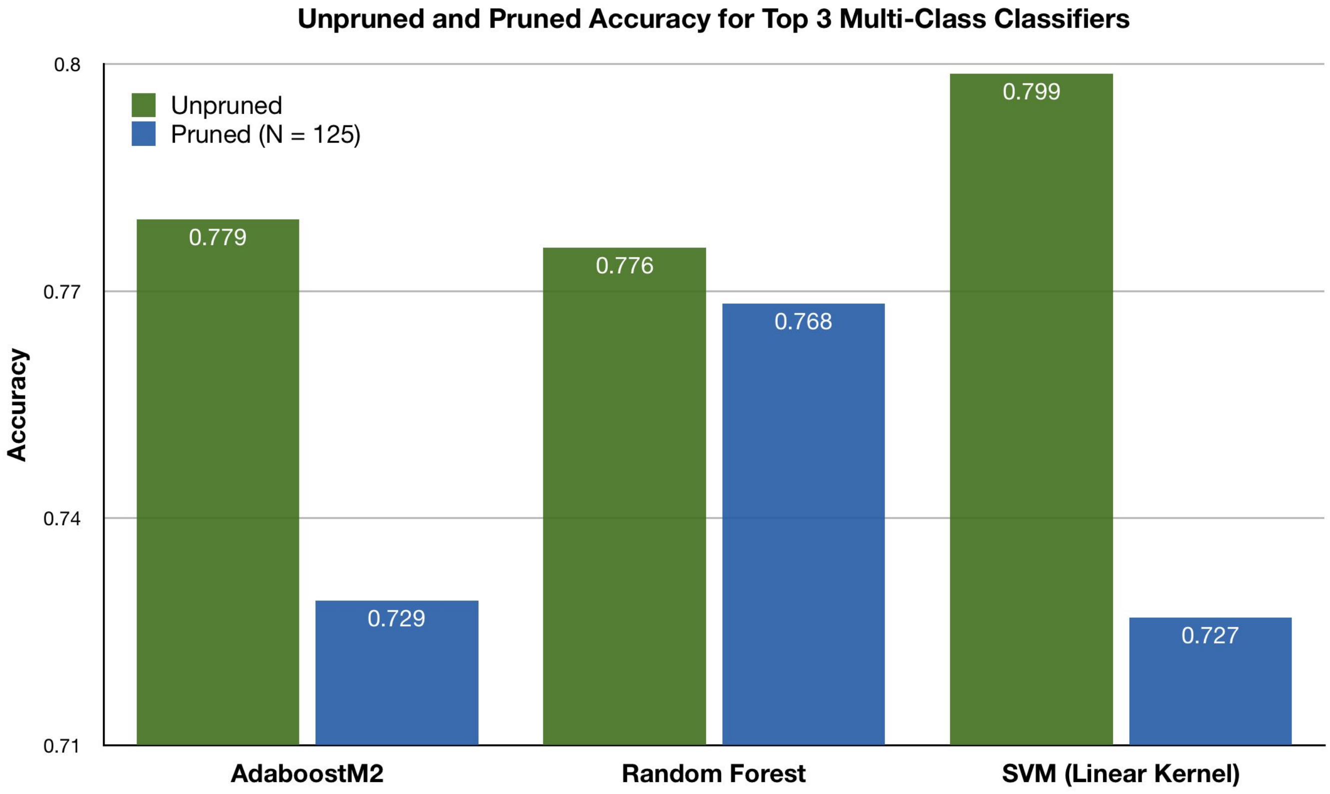
- Top 3 multi-class
- Unpruned (N = 5462 features)
- Pruned (N = 125)

Results:

- Compare unpruned & pruned accuracy

Conclusion:

- Use multi-class RF for most robust cross-validation accuracy



Analysis and Results: Frequency

Setup:

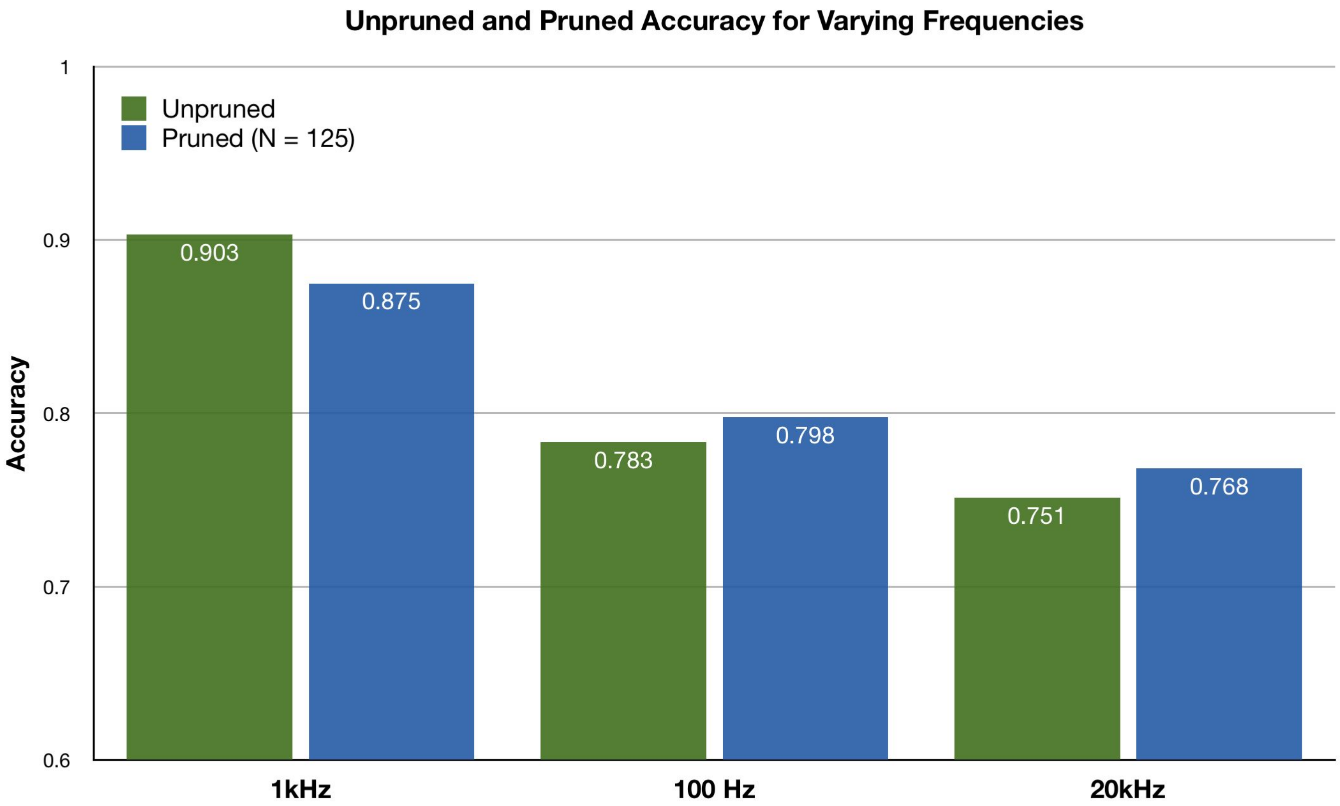
- 3 frequencies
- Unpruned
- Pruned (N = 125)

Results:

- Compare unpruned & pruned accuracy

Conclusion:

- Use 1 kHz
- 100 Hz and 20 kHz usable



Analysis and Results: Movement

Setup:

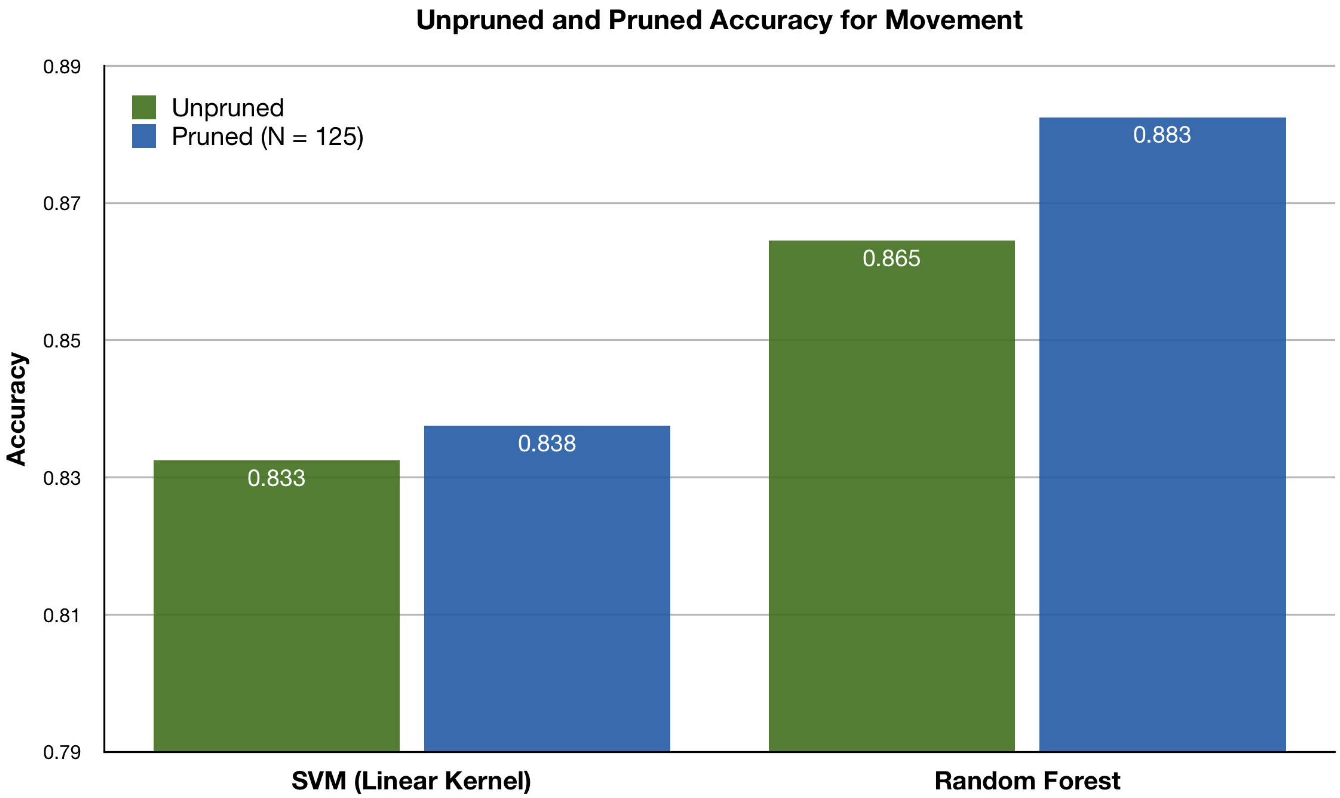
- Multi-class SVM and RF classifiers
- Unpruned
- Pruned (N = 125)

Results:

- Compare unpruned & pruned accuracy

Conclusion:

- Accuracy doesn't seem bad, but...



Analysis and Results: Movement Confusion Matrix

		Predicted class		
		Stationary (Label 0)	CCW (label 1)	CW (label 2)
True class	Stationary (label 0)	200	0	0
	CCW (label 1)	0	78	22
	CW (label 2)	0	30	70

- Confusion matrix shows that only moving data was misclassified
- Classification accuracy of just moving data is 74%, not as good
- Need more data!

Analysis and Results: Cost Matrix (1)

- Worst case scenario
  - Predict human present (predicted label = 1)
  - Room actually empty (true label = 0)
- Cost matrix
  - Row → True class
  - Column → Predicted class
  - Default: *a* = 1, *b* = 1
- Weighted misclassification, given worst case
  - Keep *b* = 1
  - Increase *a* = {1, 2, 3, 4, 5}

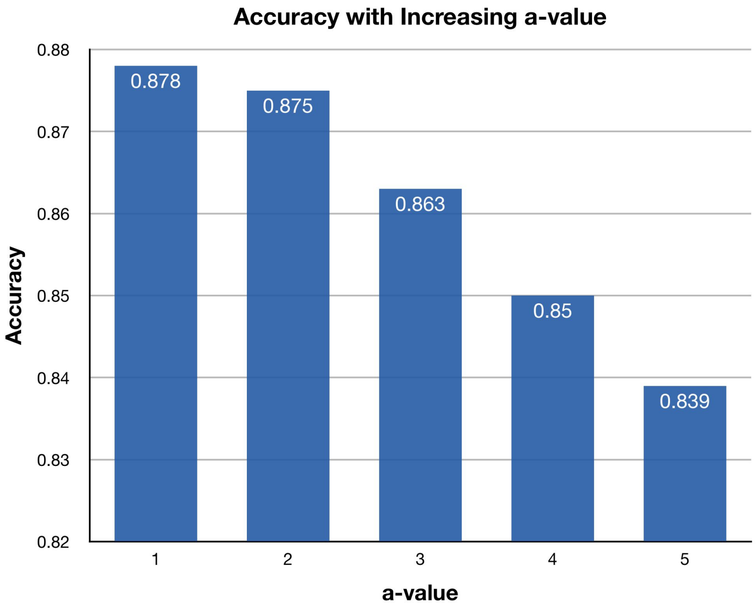
$$Cost = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}$$

	a = 1	Predicted class	
		Label 0	Label 1
True class	Label 0	348	103
	Label 1	69	831

Confusion matrix for a = 1, b = 1 (default)

Analysis and Results: Cost Matrix (2)

- Conclusion
  - Use *a* = 3 to minimize both worst-case misclassification and accuracy loss



a = 1		Predicted class	
		Label 0	Label 1
True class	Label 0	348	103
	Label 1	69	831

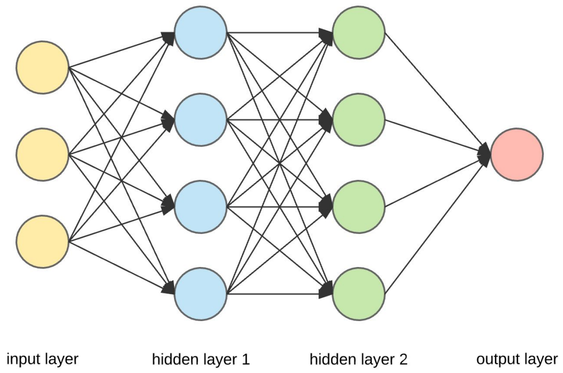
a = 2		Predicted class	
		Label 0	Label 1
True class	Label 0	382	69
	Label 1	106	794

a = 3		Predicted Class	
		Label 0	Label 1
True Class	Label 0	392	59
	Label 1	138	762

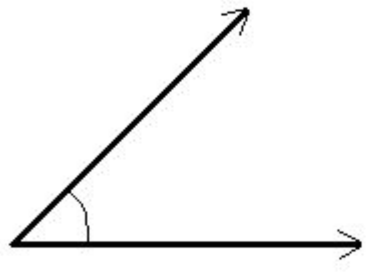
a = 4		Predicted class	
		Label 0	Label 1
True class	Label 0	403	43
	Label 1	154	746

a = 5		Predicted Class	
		Label 0	Label 1
True Class	Label 0	407	44
	Label 1	187	713

Future Directions



- Improvements to multi-class classifier
  - More data
  - More effective classifier (e.g. neural network)
  - More features
  - Better hardware
  - Different frequencies (e.g. ultrasonic)
- Extension of multi-class classifier
  - More classes, currently detects quadrants (i.e. 90 degree slices)
  - Extend to octets (i.e. 45 degree slices)
  - Exact angle via regression problem
- Protect against attacks using angle



References

- Image References
  - Amazon Echo: https://www.bhphotovideo.com/images/images2500x2500/amazon_echo_2nd_generation_charcoal_1365629.jpg
  - Google Home: https://pisces.bbystatic.com/image2/BestBuy_US/images/products/5578/5578849cv1d.jpg
  - Living Room: https://www.marniegoodfriend.com/wp-content/uploads/2018/08/Simple-Living-Room-Ideas-Awesome.jpg
  - Stick Figure: https://www.sccpre.cat/mypng/full/67-675869_stick-figure-red-man-isolated-png-image-stick.png
  - Audio fingerprint: https://images.theconversation.com/files/133561/original/image-20160809-18037-130av7l.jpg?ixlib=rb-1.1.0&q=45&auto=format&w=496&fit=clip
  - Speakers: https://www.accessories4less.com/mas_assets/cache/image/3/4/3/c/13372.Jpg
  - Neural network: https://cdn-images-1.medium.com/max/1600/1*Gh5PS4R_A5drl5ebd_gNrg@2x.png
  - Angle: https://www.analyzemath.com/Geometry/angle_1.gif
- Other References

  - [1] Amr Alanwar, Bharathan Balaji, Yuan Tian, Shuo Yang, and Mani Srivastava. 2017. EchoSafe: Sonar-based Verifiable Interaction with Intelligent Digital Agents. In Proceedings of SafeThings’17, Delft, Netherlands, November 5, 2017, 6 pages.
  - [2] Arnab Poddar, Md Sahidullah, and Goutam Saha. 2018. Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics 7, 2 (2018), 91–101. DOI:http://dx.doi.org/10.1049/iet-bmt.2017.0065

Thank you!