pindrop®

# Detection of calls from smart speaker devices

Vinay Maddali
David Looney
Kailash Patil
*Pindrop*

pindrop

# Content

- Smart speakers and detection – Motivation
- Features motivation
- LPC features
- Spectral features
- Final features
- Dataset
- Experiments
- Results
- Conclusions

# Smart Speakers

- Enabled with voice assistant technology – Amazon Alexa, Google Assistant, Siri, etc.
- Available in various makes, models and sizes.
- Year-on-year rise in use of these devices.
- Substitute smart phones for various tasks. E.g. playing music, videos, making calls.


https://www.commonsense.org/education/articles/compare-the-privacy-practices-of-the-most-popular-smart-speakers-with-virtual-assistants

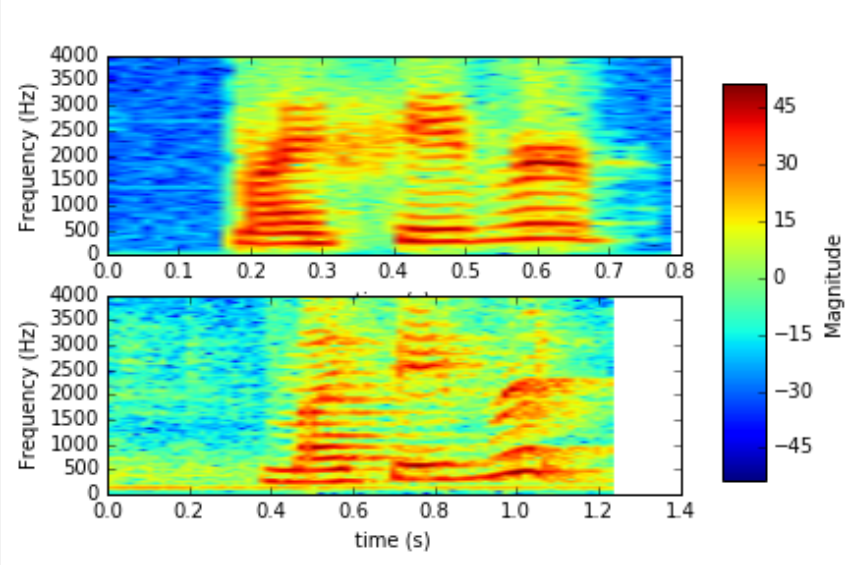
## Why is detection important?

Smart speaker devices make Voice over Internet Protocol (VoIP) calls to the Public Switched Telephone Network (PSTN) retaining the user's cell or landline number.

Systems that relate audio attributes to the metadata will be affected.

- Authentication systems in call centers
- Calibration in speaker identification systems
- Audio forensics.

# Features motivation




- Shows the spectrograms of the same utterance by the same speaker from cell (top) and smart speaker (below).
- Calls from smart speakers are different due to various factors: number of microphones, microphone array configuration, type of beamforming and denoising algorithms.

**We observe a loss in harmonic structure, sharpness of onsets and channel differences.**

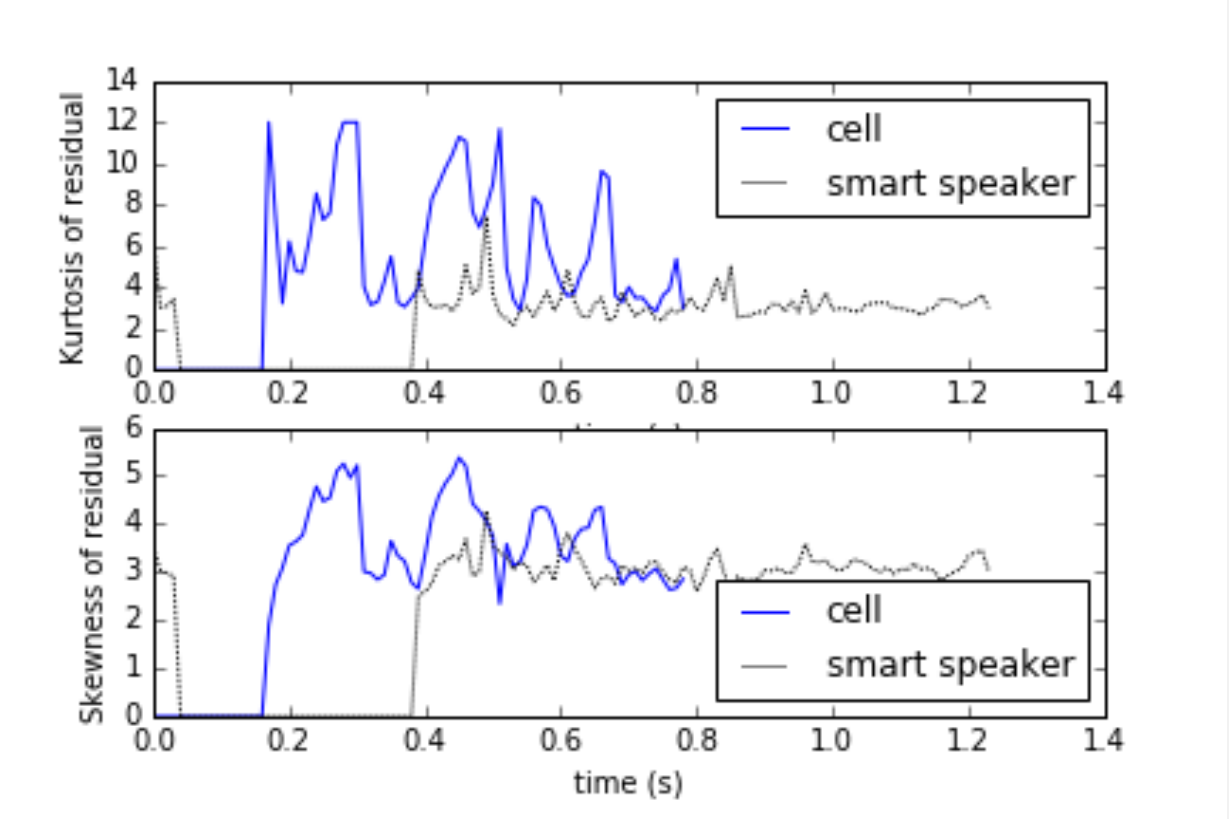
# LPC features

## Background

- LPC prediction residual of clean speech has true excitation peaks of similar magnitude.
- Become erroneous and larger in magnitude with increasing reverberation [1].

**Smart speakers capture more reverberant speech than cell phones due to the usage patterns.**

**Higher kurtosis and skewness values for the same utterance as shown in Figure.**




[1] Patrick A. Naylor and Nikolay D. Gaubitch. Speech Dereverberation. Springer, 2010.

# Spectral features

- In order to differentiate based on noise, network and channel differences, we used the following from the librosa library [2]:
  - Spectral flatness
  - Spectral rolloff
  - Spectral centroid
  - Spectral bandwidth
- We developed a harmonic mean Fast Fourier Transform (FFT) measure which computes a harmonic mean across time for each frequency bin of the short-time Fourier transform (STFT) outputs:

$$h_k = \frac{T}{\sum_{t=1}^{T} \frac{1}{S(k,t)}}$$

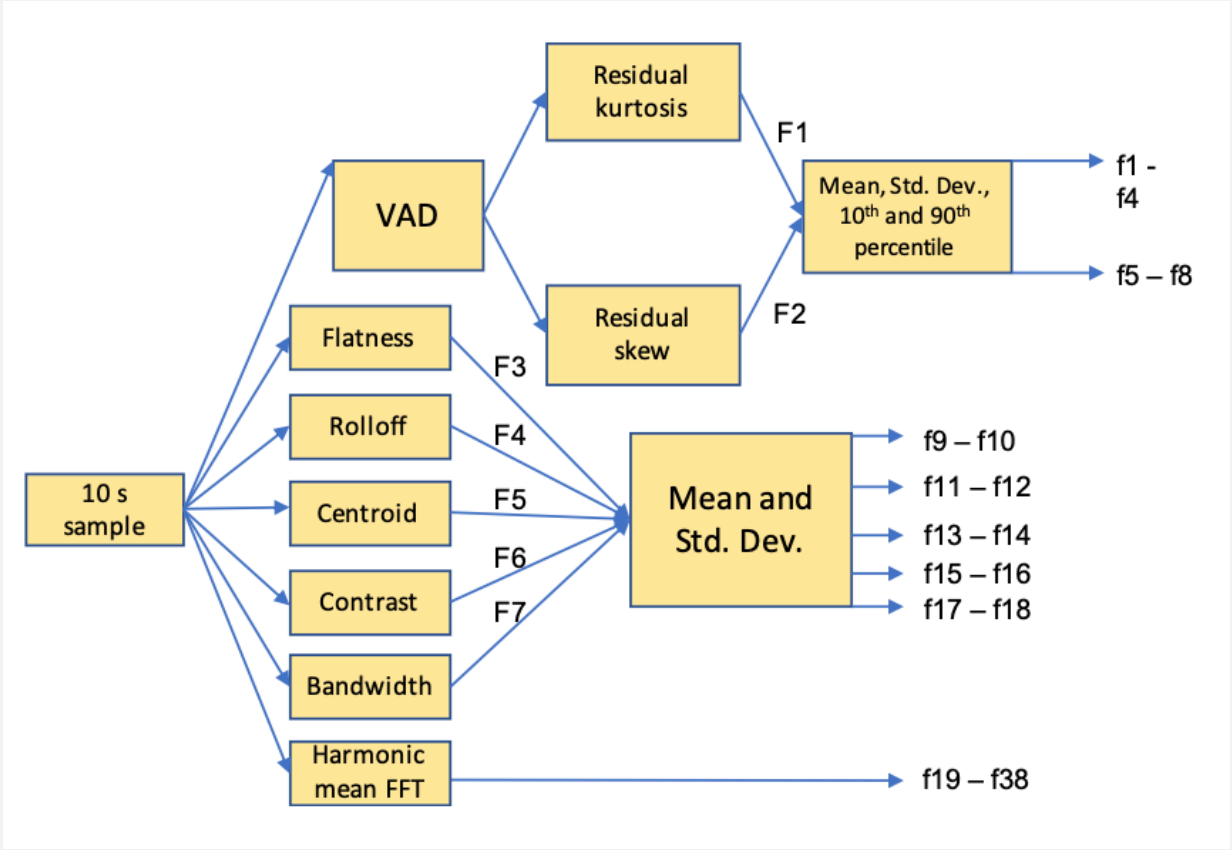
where S(k,t) is the spectrogram across frequency f and time t and T is the number of time frames.

[2] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. *librosa: Audio and music signal analysis in Python*. Proceedings of the 14th Python in Science Conference, 2015.

# Features – more details

- **LPC features:**
  - Computed on 20 ms voiced frames with 50% overlap.
- **Harmonic mean:**
  - 128 pt FFT and 50% overlap.
- **Other spectral features:**
  - 2048 pt with 50% overlap.

Mean, std. dev., 10th and 90th percentile were computed for LPC residual features and some spectral features.



pindrop

# Dataset

- Actual phone calls collected through crowdsourcing.
- Speakers were asked to speak for 60s without scripted speech to emulate natural speech.
- Each participant made 4 calls : 2 from cell and 2 from a smart speaker.
- Recorded attributes:
  - make and model of the smart speaker.
  - distance from smart speaker
  - mode-of-usage of cell phone.
- Collected a total of 552 calls from 138 users and audio was sampled at 8kHz.

pindrop

# Dataset

TABLE I
MODELS OF SMART SPEAKERS AND THEIR USER COUNTS.

| Smart speaker models | User count |
|---|---|
| Amazon Echo (AEcho) | 42 |
| Amazon Echo Dot (AEDot) | 41 |
| Amazon Echo Show (AEShow) | 4 |
| Amazon Echo Spot (AESpot) | 4 |
| Amazon Echo Plus (AEPlus) | 3 |
| Google Home (GHome) | 25 |
| Google Home Mini (GHmini) | 19 |

TABLE II
MODES OF USAGE AND SPEAKER DISTANCE CATEGORIES WITH THEIR USER COUNTS

| Cell phone usage mode | User count |
|---|---|
| Normal(to-ear) | 70 |
| Speakerphone | 55 |
| Earphones | 6 |
| Bluetooth Headset | 7 |
| **Smart speaker distance** | **User count** |
| <10 cm | 42 |
| 10 cm-100 cm | 77 |
| 100 cm-200 cm | 31 |
| >200 cm | 8 |

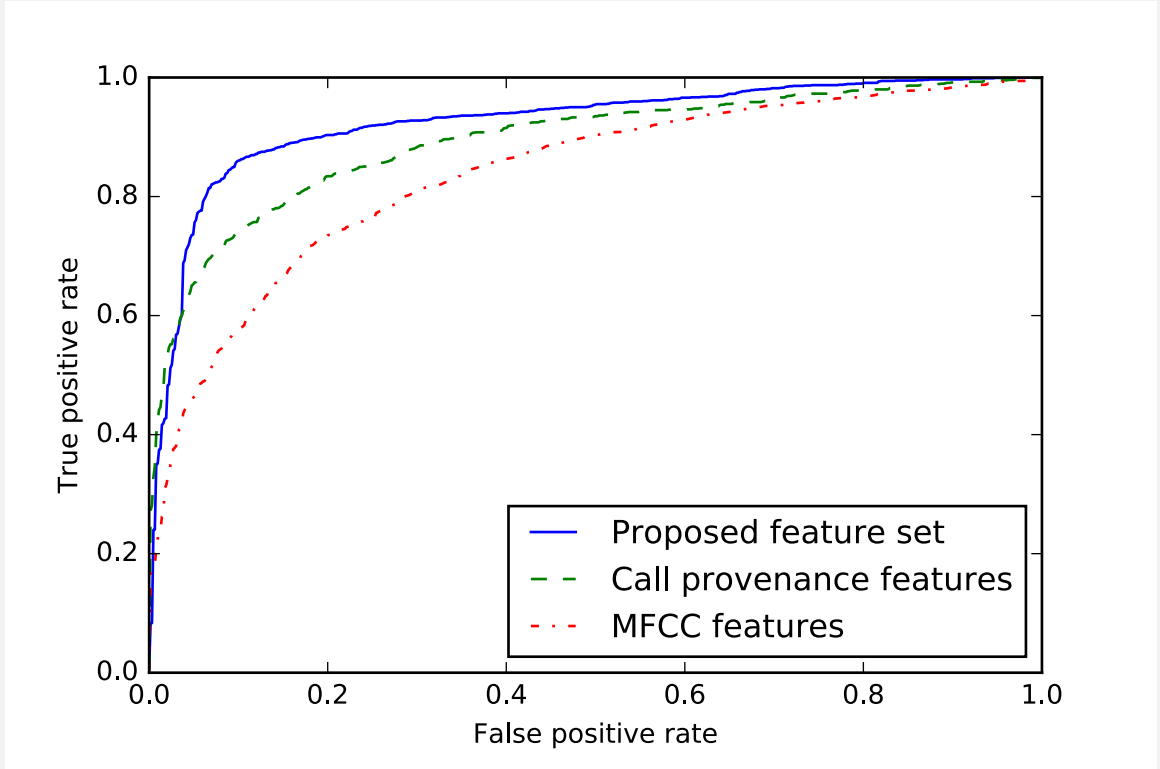
# Experiments

- Calls were divided into 10 s samples.
- Features f1-f38 were used.
- Performance was assessed using five-fold cross validation with 80-20 train-test split with no overlap in speakers.
- Best performance for each from SVM, kernel SVM, gradient boosting, random forest, AdaBoost and logistic regression classifiers was used.
- Baseline of Call provenance [3] features and MFCC features was used to compare performance of this system.

[3] Vijay A. Balasubramaniyan, Aamir Poonawalla, Mustaque Ahamad, Michael T. Hunter, and Patrick Traynor. *Pindr0p: Using single-ended audio features to determine call provenance*. CCS 10 Proceedings of the 17th ACM conference on Computer and communications security, 2010, pp. 109–120.

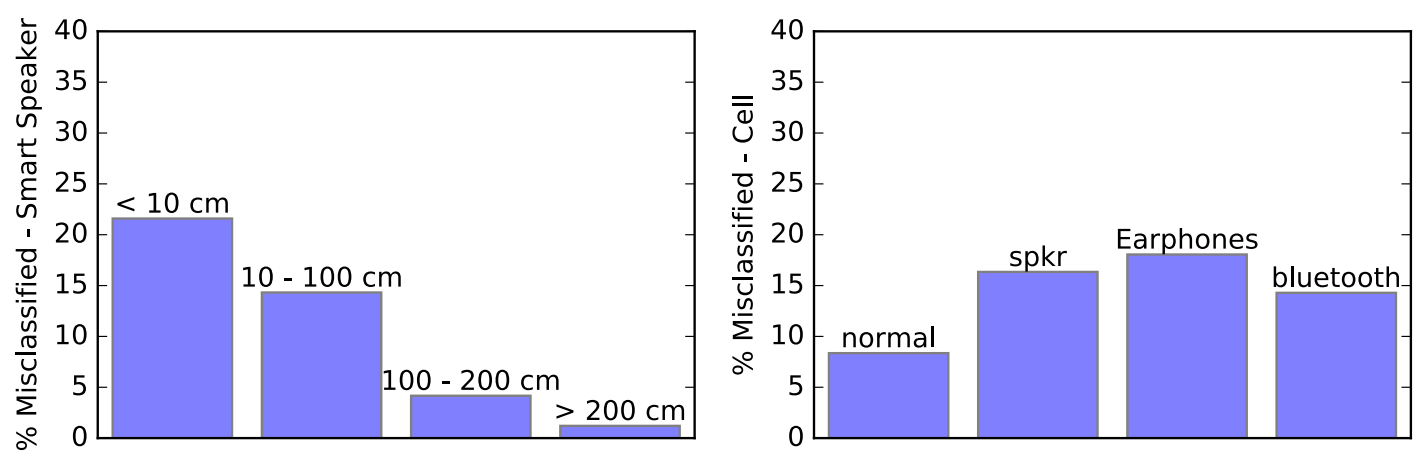
# Results - ROC

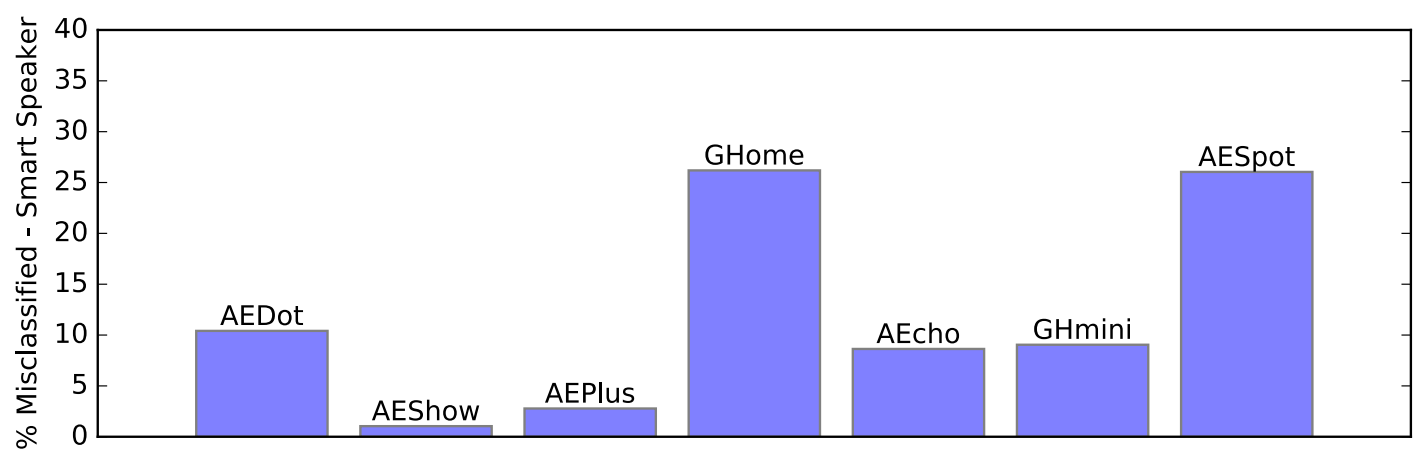
| System | EER |
|---|---|
| Proposed | 12.6% |
| Call provenance | 17.6% |
| MFCC | 23.8% |




**Proposed system shows significant improvement over baseline!**

# Misclassification analysis




- Performance for cell is best in normal mode as this has least reverberation and noise.
- Performance for smart speakers is best when used from far away, >100 cm.
- Worst performance in all cases was 21%.

**Features are not overly biased on mode-of-usage for either devices.**

# Conclusions

- Proposed an approach to differentiate smart speaker calls from cell phone calls.
- Proposed audio feature set to detect differences in reverberation, noise and spectral characteristics.
- A dataset was collected through crowdsourcing with participants using both devices in different modes.
- Proposed system differs from previous works as it detects smart speaker calls and uses a dataset that closely resembles real-world call scenarios.
- Improvement over baseline systems, call provenance and MFCCs by 28% and 47% respectively.

pindrop®

# THANK YOU