Audio-Based SPAM Call Detection

Benjamin Elizalde, Dimitra Emmanouilidou

benjaminm@microsoft.com, dimitra.emmanouilidou@microsoft.com

Motivation and Problem

- SPAM calls are organized attempts with the purpose of marketing, spreading unwanted information, and scamming.
- The US is among the most spammed countries in 2020, with **28 calls per month per person**.
- The US had **46 billion Robocalls** in 2020.

Background

SPAM call detection have seen multiple approaches, **but are not enough**.

- Tracking Call Detail Records (call origin, phone number, call duration) are effective, but new unseen records come every day.
- Analyzing the generated call transcript, but intruding users privacy.

Solution

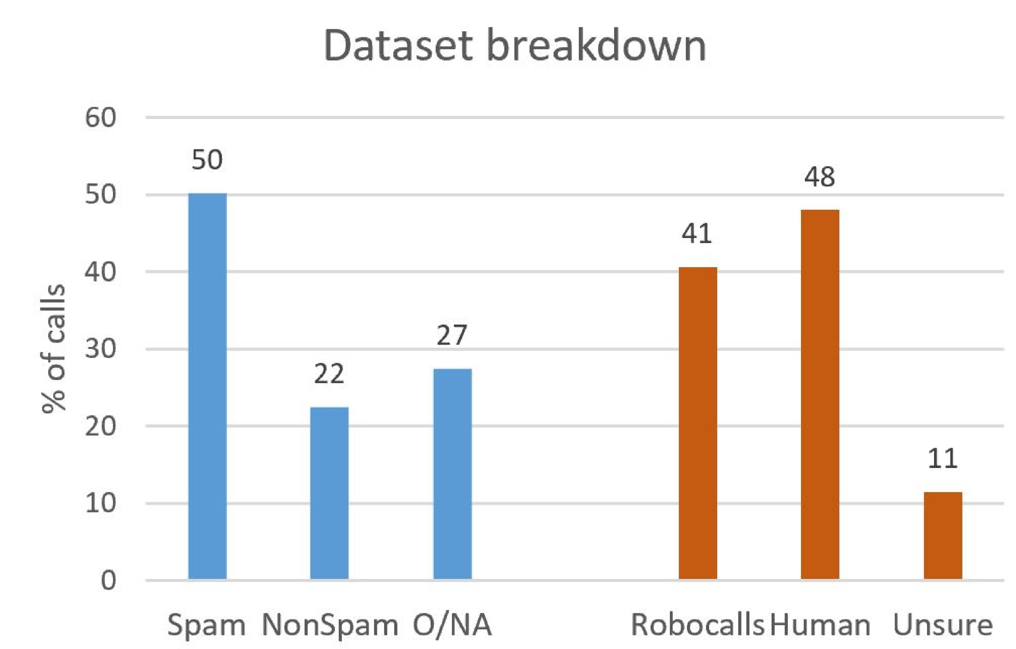
- We proposed audio-based SPAM detection for voicemail recordings.
- Audio-content analysis preserves privacy because it does not look into the spoken content or transcripts.

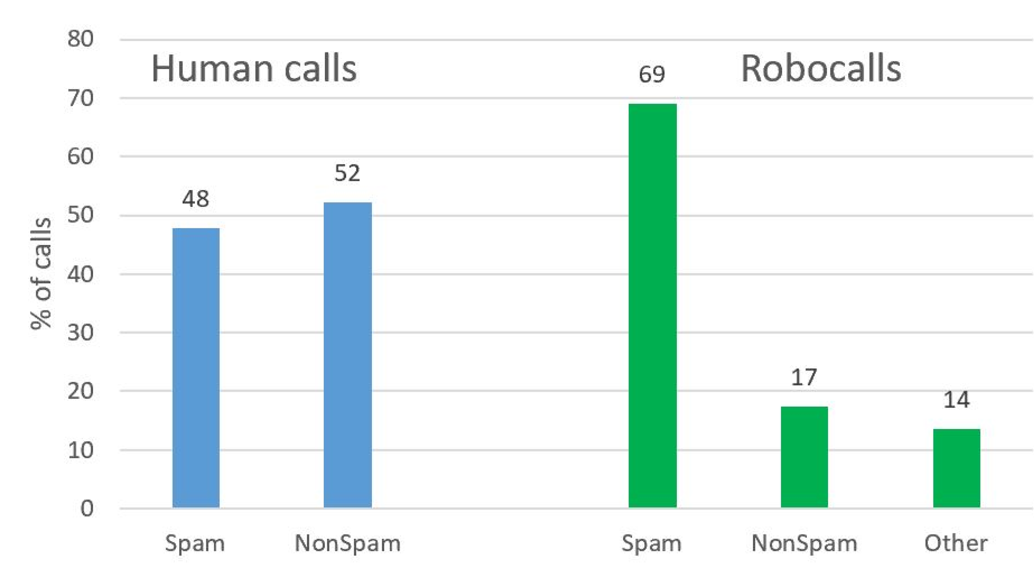
Dataset

- Collected 596 voicemails from different users with median duration of **30 secs** ± 25 secs.
- Data annotated as {**Human vs Robocalls**} and {**SPAM vs Non-SPAM**}.
- **6.3 ± 3.4 secs** was the time it took the annotators to decide if a voicemail was SPAM.
- Among all Non-SPAM calls, **90% were Human calls** and 10% were Robocalls.
- Among all SPAM calls, **39% were Human calls** and 56% were Robocalls.
- In Human calls, the ratio of **female** speakers was **2:1 for SPAM** and 4:3 for Non-SPAM.

Are Robocalls the real offenders?

- Not uniquely. Robocalls made up for 34% of SPAM calls in the US in 2019.





Can we identify Human calls from Robocalls using acoustic features?

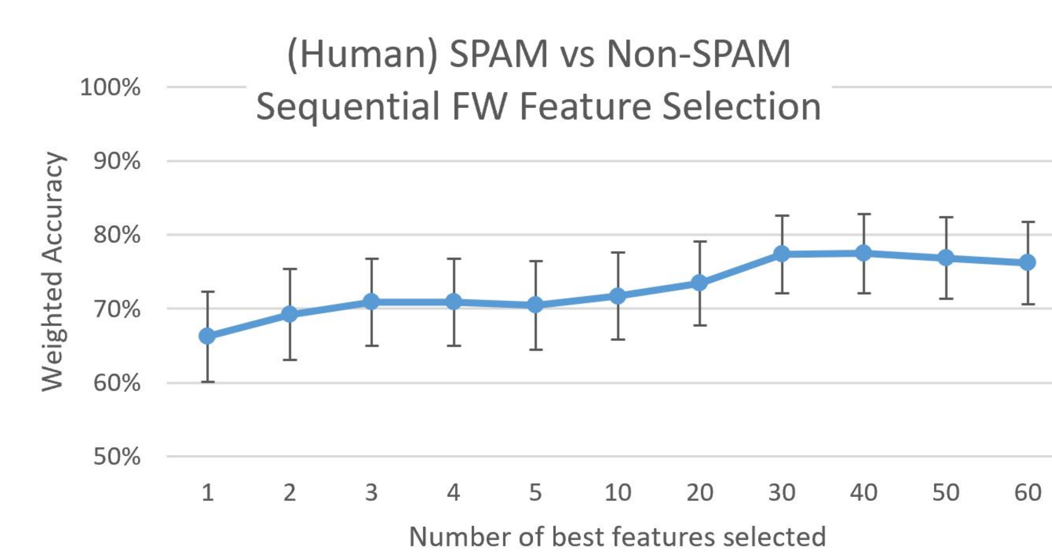
	Accuracy
Human Vs Robocalls (K-SVM)	93.12 (±2.33) %

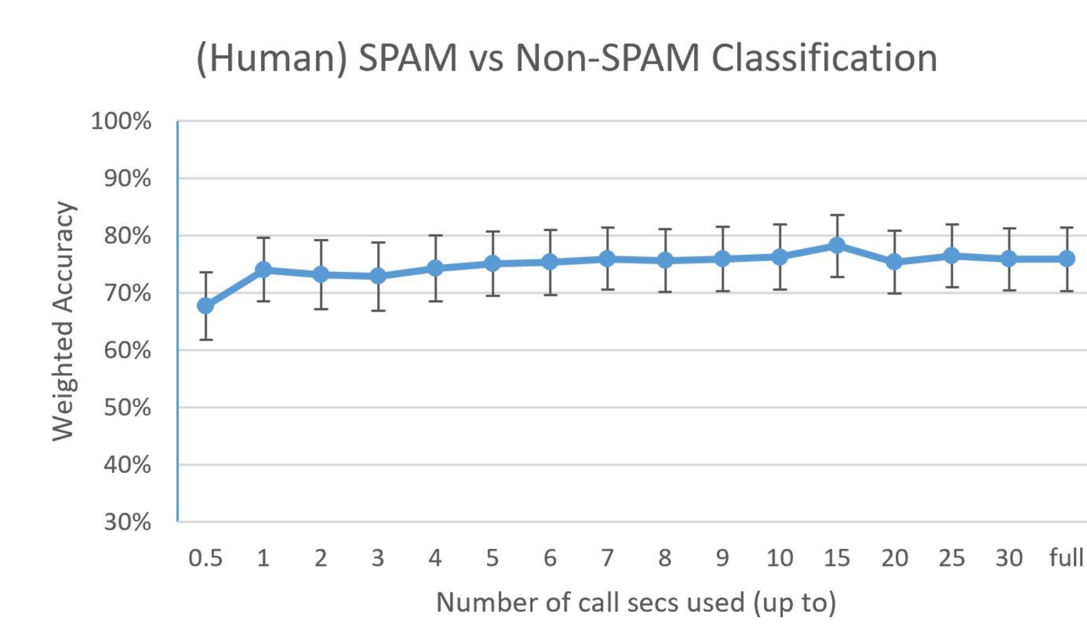
- Features: Opensmile's GeMAPSv01b spectrotemporal statistics, 62 dim. Classification: Binary rbf-SVM, 80-20% split, 500 M.C. runs.
- 79% accuracy using a SINGLE feature, 88% using best five. Top feature selection:
  - $VoicedSpecSlope0\text{-}500V_{\mu}$ (spectral)
  - $F0FallingSlope_{\sigma}$, $F0RisingSlope_{\sigma}$ (spectral),
  - $F0_{Perc20}$, $F0_{Perc80}$ (freq)
  - $UnvoicedSegmLength_{\sigma}$ (temporal)
  - $PerceivedLoudness_{perc50}$ (energy)

For human callers, can we identify SPAM calls?

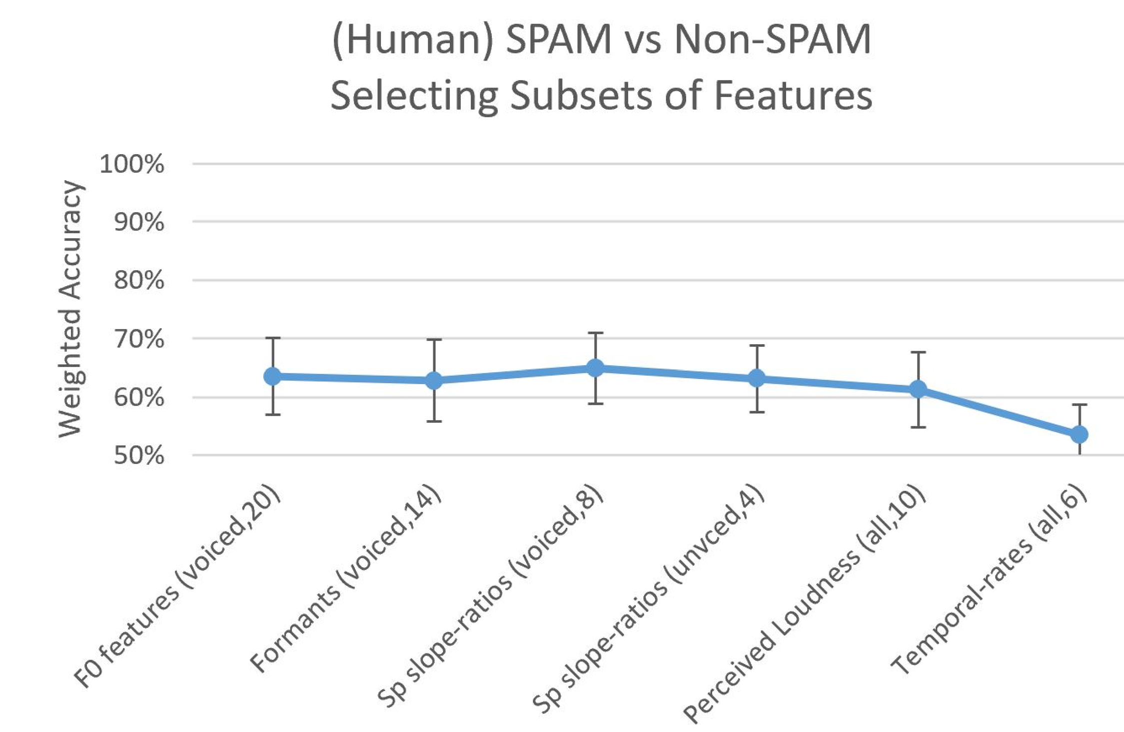
	Accuracy
SPAM vs Non-SPAM (K-SVM)	75.86 (±5.51) %
SPAM vs Non-SPAM (CNN)	82.60 (±4.76) %

- (K-SVM) Features: Opensmile's GeMAPSv01b spectrotemporal statistics, 62 dim. Classification: Binary rbf-SVM, 80-20% split, 500 M.C. runs.
- (CNN) Features: LogMel spectrogram 32 channels, 32msec frames. Classification: 2-block CNN (32,5,1|64,3,2), 75-10-15% split.
- For (K-SVM), 65% accuracy using SINGLE feature; 70% using five. Top feature selection:
  - $UnVoicedSpecSlope0\text{-}500V_{\mu}$ (spectral)
  - $Voicedsegm/sec$ (temporal)
  - $F2_{\mu}$, $F0FallingSlope_{\sigma}$ (freq)
  - $VoicedSpecSlope0\text{-}500V_{\sigma}$ (spectral)
  - $Harmonic\text{-}NoiseRatio\mu$ (energy)

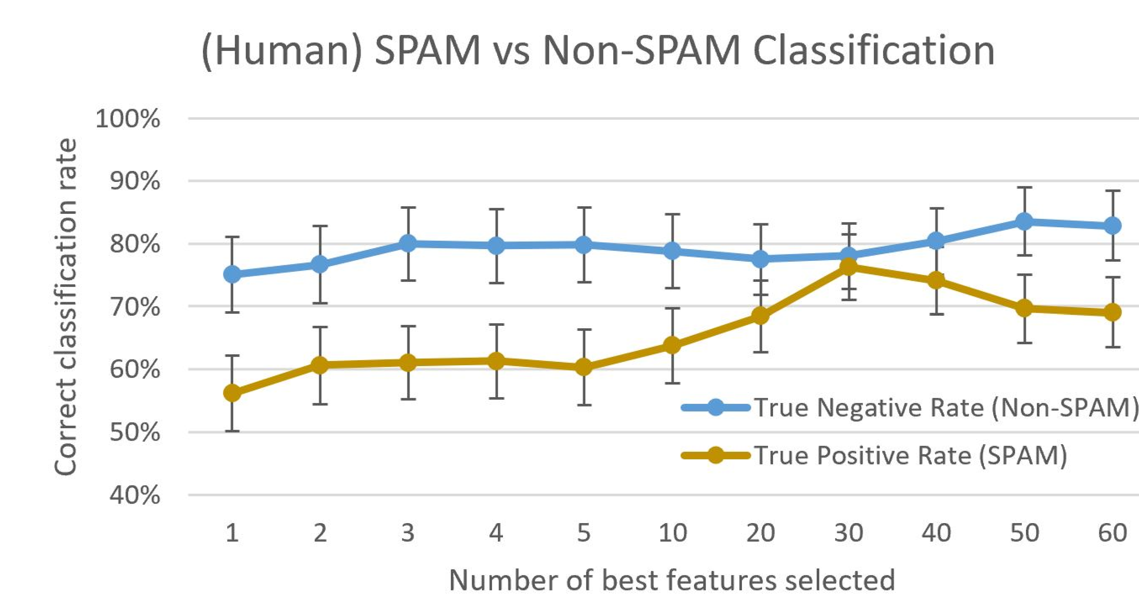




- We grouped features into subset types to compare their contribution to Human SPAM classification.
- The unvoiced feature subset contributes as much as other voiced feature subsets.



- A small number of features performs better at identifying True Negative cases (Non-SPAM) rather than True Positives (SPAM). A larger number of features is needed to boost up True Positive (SPAM) classification rate.



Conclusion

- We found that audio content in voicemails can be used to distinguish SPAM vs Non-SPAM (85% acc), Robocall vs Human calls (93% acc).
- Robocalls made up 56% of our dataset; we can identify them well with a few features.
- A large number of human voicemails were labeled SPAM (48%), which underlines the need to process calls beyond Robocalls.
- Unvoiced regions features are useful for classifying Human SPAM vs Non-SPAM.
- Even for human calls, just a few secs are enough to distinguish SPAM from non-SPAM.