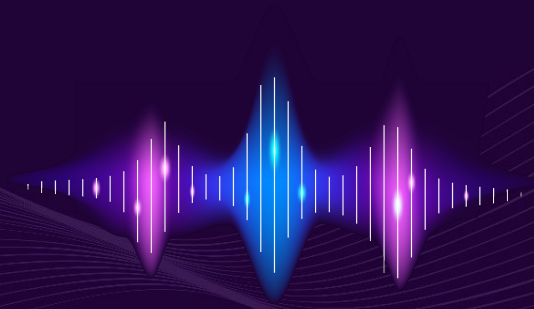
LEHIGH UNIVERSITY

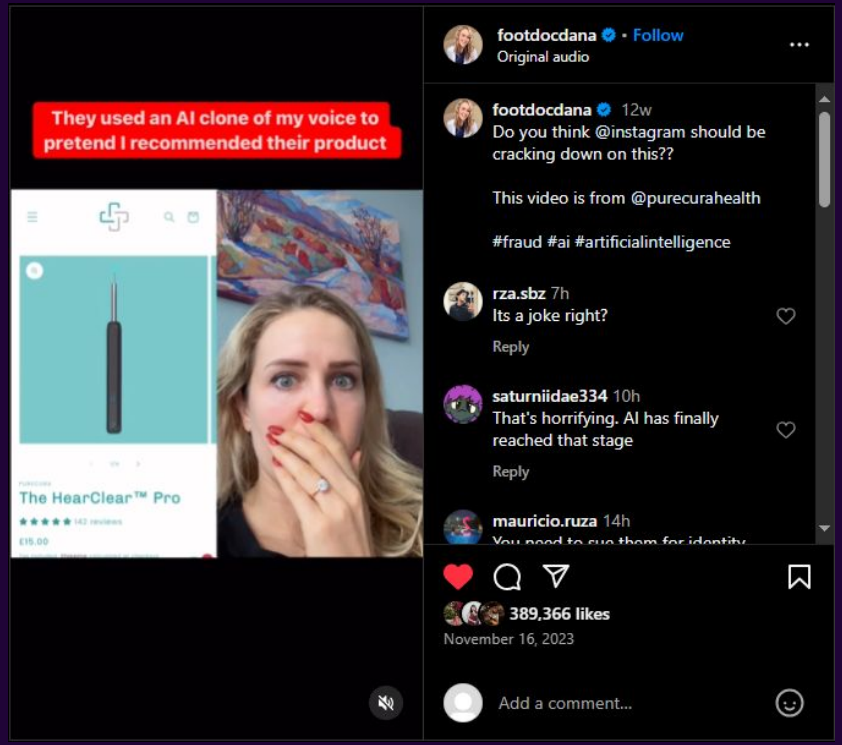
Voice Cloning Detection: Deep4SNet Counter to SiF-DeepVC

Thomas Snipes, MS. CSE
Shuang Lin, BS. COGS



Impersonation

Doctor's voice cloned to advertise a product as "doctor approved"

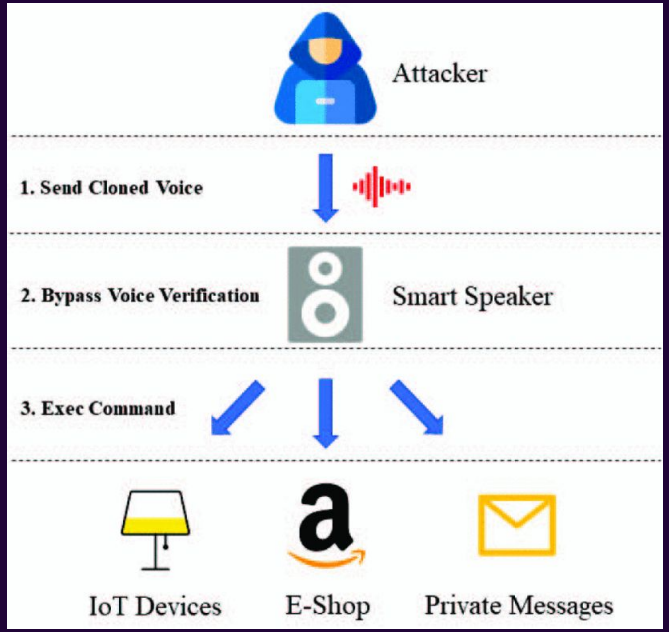


@footdocdana

LEHIGH UNIVERSITY

Device Hijack

Smart speaker hijacking to perform personal actions



AI-Synthesized Voice Threat Model in Home IoT - X. Liu et al.

LEHIGH UNIVERSITY

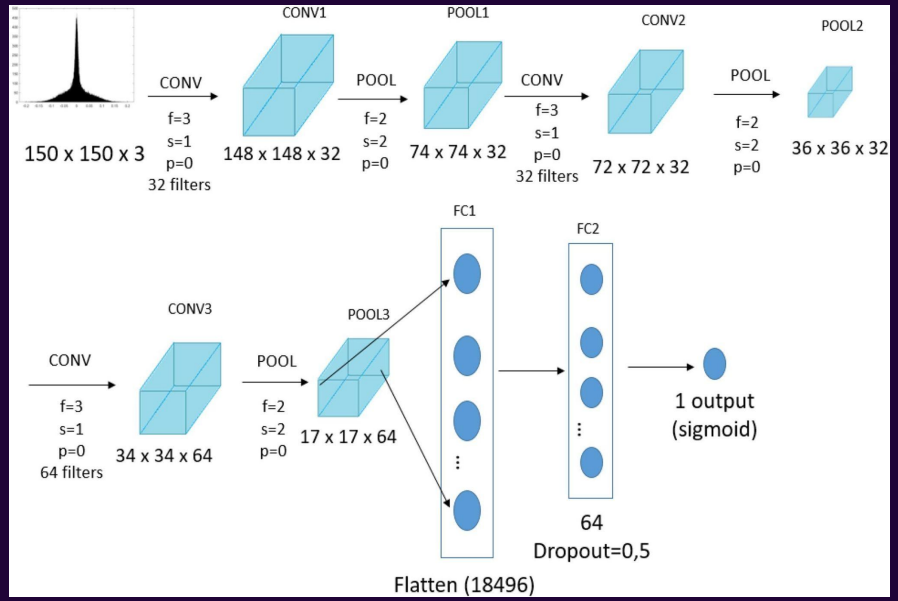
Goal

Improve upon the Deep4SNet deep learning model to counter the voice clones camouflaged by SiF-DeepVC

LEHIGH UNIVERSITY

Deep4SNet

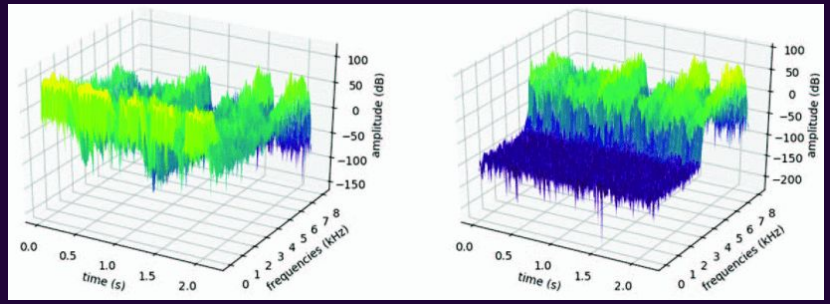
- Convolutional Neural Network
- Trained on Deep Voice & Imitation
- 98.5% global accuracy
- Computer vision approach
  - Histogram input
  - Supervised learning
  - Less layers



LEHIGH UNIVERSITY

SiF-DeepVC vs. Deep4SNet

- 300 Hz to 3400 Hz = human voice
- 4000 Hz+ = vast majority of SiFs
- Can effectively confuse the detection systems.



- False Positive Percentages
- High accuracy ≠ correctness
- Deep4SNet = 66.61%

Approach	Type	Language	Baseline FPR	Denoised FPR	Diff *
Farid et al.	CV-based	English	67.70%	75.09%	↑ 10.92%
		Mandarin	45.39%	84.37%	↑ 85.88%
Deep4SNet	CV-based	English	66.61%	59.85%	↓ 10.15%
		Mandarin	90.95%	99.37%	↑ 9.26%
DeepSonar	NNF-based	English	52.56%	53.43%	↑ 1.66%
		Mandarin	37.99%	36.68%	↓ 3.45%
RawNet2	E2E-based	English	94.43%	97.22%	↑ 2.95%
		Mandarin	47.70%	55.74%	↑ 16.86%

* Compared with original baseline results

Data - 1/2

SiF-DeepVC

- .WAV Audio files
- Fake - 4.5k samples
  - 1k Target set
- Real - 4.5k samples

Source	Files	Label
FoR	4500	Real
Farid et al.	2995	Fake
RawNet2	1505	Fake
Deep4SNet	1000	Fake

TABLE I: Audio Files From SiF-DeepVC

Split

- Train: 70%
- Validation: 15%
- Test: 15%

Data - 2/2

H-Voice

- .jpg Histograms
- Using all files
- Fake - 3404 samples
- Real - 3268 samples

Source	Files	Label
Original	3268	Real
Imitation	3260	Fake
Deep Voice	144	Fake

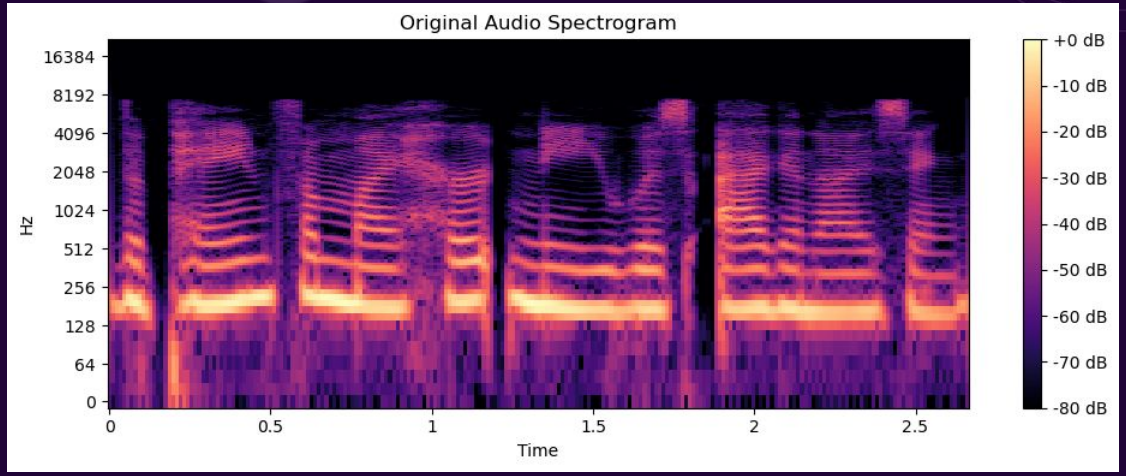
TABLE II: Audio Files From H-Voice

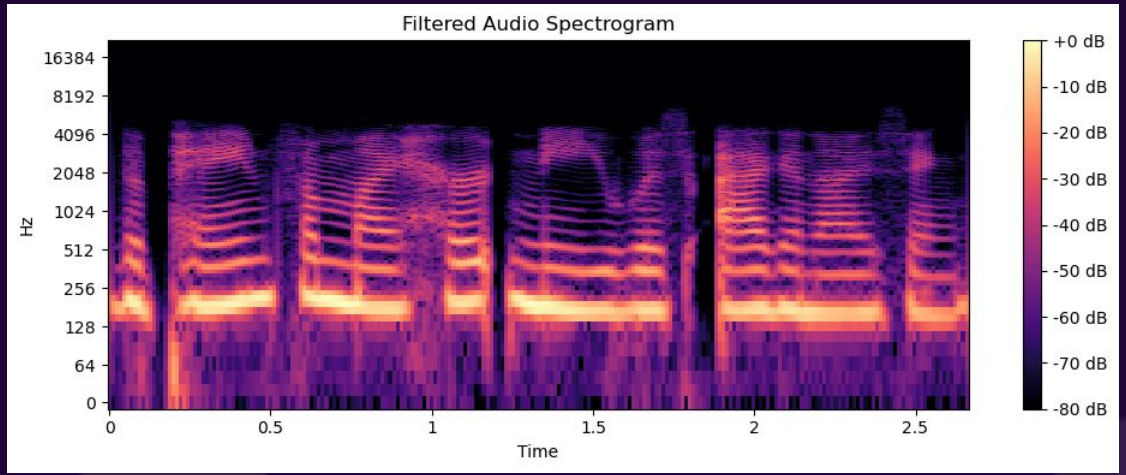
Split (provided)

- Train: 61.57%
- Validation: 25.90%
- Test: 12.53%

LEHIGH UNIVERSITY

Filtering





LEHIGH UNIVERSITY

H-Voice

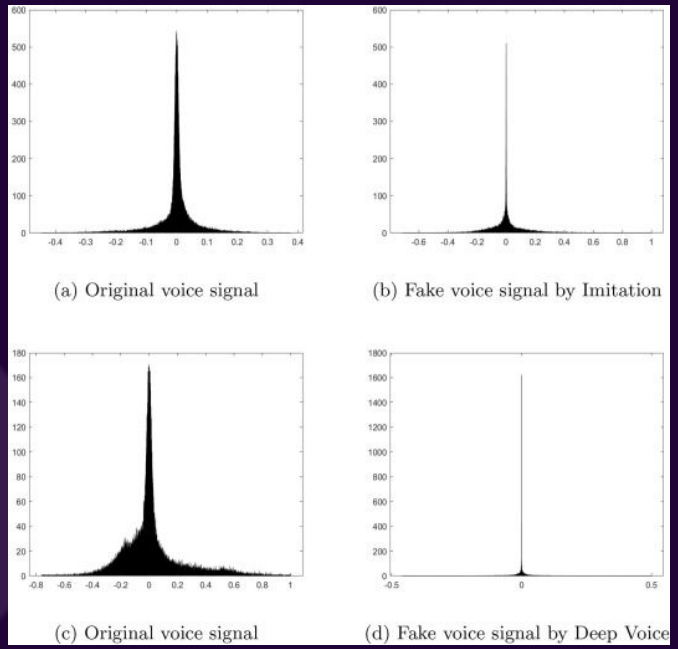


Figure 8 - Deep4SNet: deep learning for fake speech classification

SiF-DeepVC (Ours)

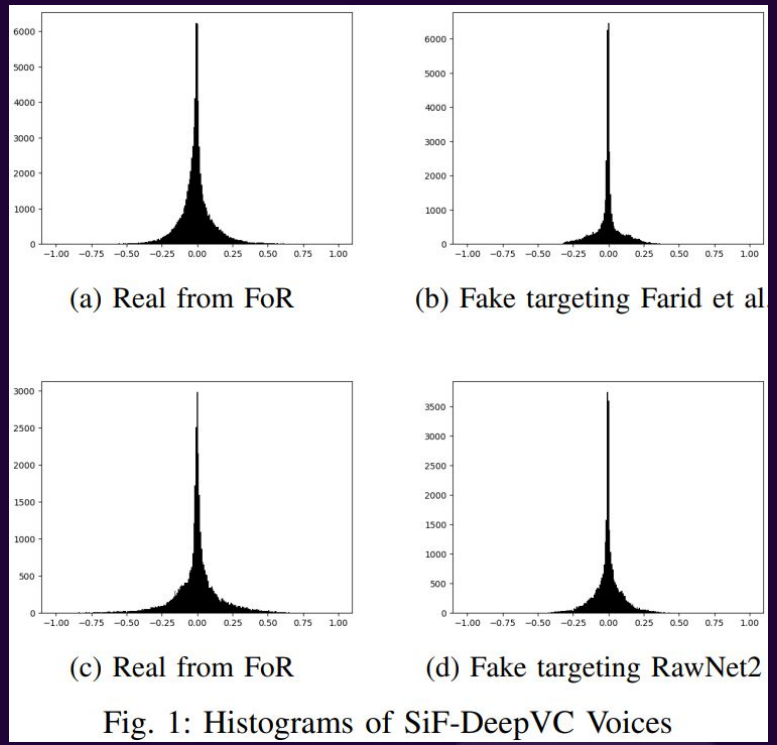


Fig. 1: Histograms of SiF-DeepVC Voices

LEHIGH UNIVERSITY

Methodology

1. Gather a large amount of diverse data
2. Create histograms from audio
3. Train models in 5 categories:
    a. SiF Regular
    b. SiF Filtered
    c. H-Voice
    d. H-Voice + SiF Regular
    e. H-Voice + SiF Filtered
4. Recreate the Deep4SNet CNN as a baseline
5. Match the accuracy and beat the FPR

LEHIGH UNIVERSITY

Algorithm

- Custom combined & filtered datasets
- Model using different architectures:
  - Multiple dropout layers: test for overfitting
  - Added convolutional layers: test overcomplexity

LEHIGH UNIVERSITY

Results - 1/3

TABLE III: Test on Data - SiF-DeepVC - Regular

Model	Accuracy - Test	FPR - Test	Accuracy - Target	TPR - Target
Original-HVoice	53.78%	52.38%	9.00%	17.71%
Our-HVoice	50.74%	52.38%	0.00%	17.71%
Our-HVoice-Dropout	50.74%	52.38%	0.00%	15.55%
Our-HVoice-Deep	49.26%	52.38%	100.00%	15.55%
Our-SiF-Regular	93.85%	53.08%	94.00%	17.71%
Our-SiF-Regular-Dropout	93.48%	52.03%	94.80%	17.71%
Our-SiF-Regular-Deep	93.93%	48.87%	98.30%	17.71%
Our-SiF-Filtered	64.52%	52.38%	62.40%	15.55%
Our-HVoice_SiF-Regular	53.11%	52.38%	99.30%	15.55%
Our-HVoice_SiF-Filtered	56.00%	52.38%	99.00%	15.55%

LEHIGH UNIVERSITY

Results - 2/3

TABLE IV: Test on Data - SiF-DeepVC - Filtered

Model	Accuracy - Test	FPR - Test	Accuracy - Target	TPR - Target
Original-HVoice	51.04%	49.81%	5.70%	21.50%
Our-HVoice	48.52%	99.86%	0.10%	21.50%
Our-HVoice-Dropout	48.59%	98.92%	0.00%	0.02%
Our-HVoice-Deep	51.41%	98.92%	100.00%	17.71%
Our-SiF-Regular	50.74%	90.06%	11.30%	0.02%
Our-SiF-Regular-Dropout	52.96%	93.66%	8.40%	0.02%
Our-SiF-Regular-Deep	50.15%	91.35%	14.00%	0.02%
Our-SiF-Filtered	94.44%	44.96%	98.30%	21.50%
Our-HVoice_SiF-Regular	67.33%	17.43%	92.60%	21.50%
Our-HVoice_SiF-Filtered	70.30%	23.20%	92.10%	21.50%

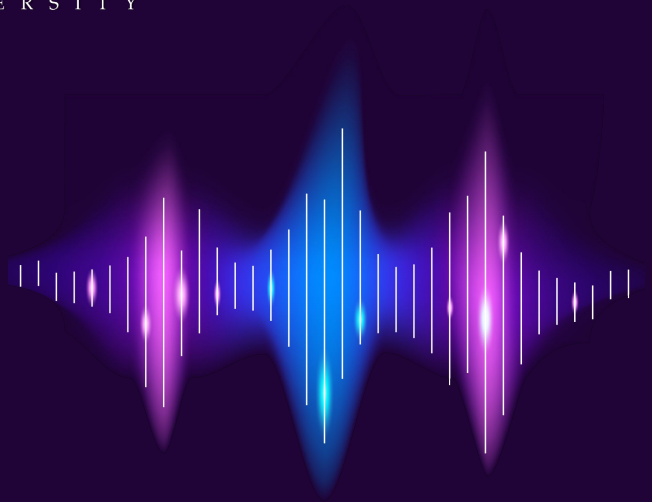
LEHIGH UNIVERSITY

Results - 3/3

TABLE V: Test on Data - H-Voice

Model	Accuracy - Test	FPR - Test
Original-HVoice	97.61%	42.92%
Our-HVoice	54.07%	47.35%
Our-HVoice-Dropout	93.18%	48.89%
Our-HVoice-Deep	54.07%	98.92%
Our-SiF-Regular	56.71%	28.54%
Our-SiF-Regular-Dropout	55.38%	20.13%
Our-SiF-Regular-Deep	54.78%	0.66%
Our-SiF-Filtered	49.28%	73.23%
Our-HVoice_SiF-Regular	95.93%	45.35%
Our-HVoice_SiF-Filtered	98.68%	44.03%

LEHIGH UNIVERSITY



Thanks!

Any Questions?