Task 11
AI based intrusion detection

SE: Detecting Realtime Deepfakes

Speaker: Dr. Yisroel Mirsky

Team:

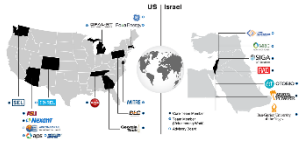
Guy Frankovits, Lior Yasur,
Fred M. Grabovski

CBG
Cyber@Ben-Gurion
University of the Negev

The Threat

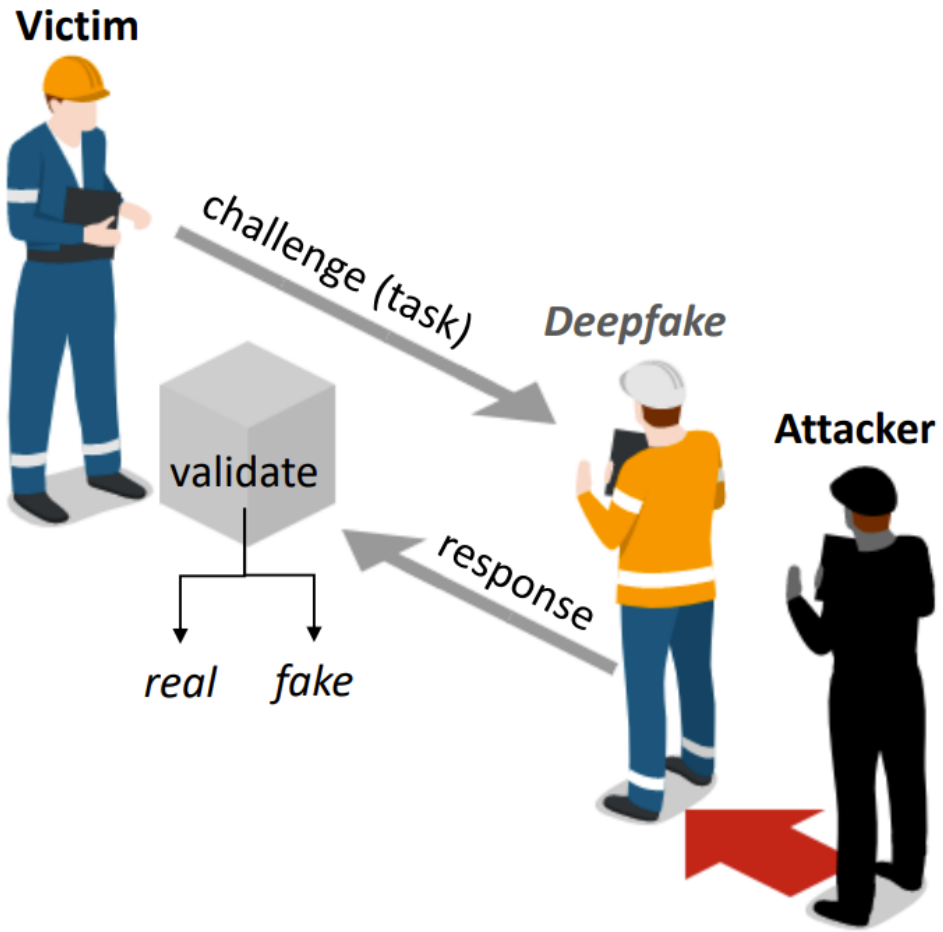
Research Plan

Present = *T+12*



Solution: DF-Captcha

A Turing Test on content generation



Research Plan

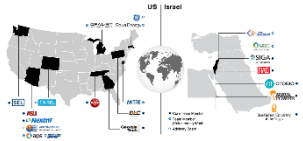
AUDIO

1. **Implement RT-DF Technologies**
   - Survey STOA Audio Cloning
   - Collect Existing Code
   - Implement Methods
   - Evaluate Quality (blind)
2. **Analyze RT-DF Limitations**
   - Stress training data limits
   - Stress tech limits
   - Stress scope limits
3. **Develop DF-Captchas (challenges)**
   - Enumerate challenges with usability
4. **Develop response analysis**
   - Static Anomaly Detection (artifacts)
   - Temporal Anomaly Detection (failure point)

VIDEO

Repeat 1-4

How It Works



RT-DF models are limited by their

Technology

1. Inference speed
2. Feature representations
3. Training

Resources

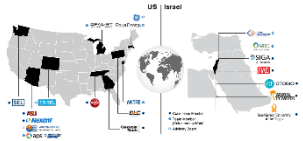
1. Data Collection
2. Knowledge
3. Labeling
4. Assets

We can <u>force</u> a RT-DF to break by pushing these limitations

Example:
Training a RT-DF model to be excellent at both speech and sing is hard.

If the caller tries to sing, the model will cause distortions in the audio.

How It Works



The Captcha System

1) $A \rightarrow B : c$
   A sends caller B a challenge
   E.g., "hum a specific song"

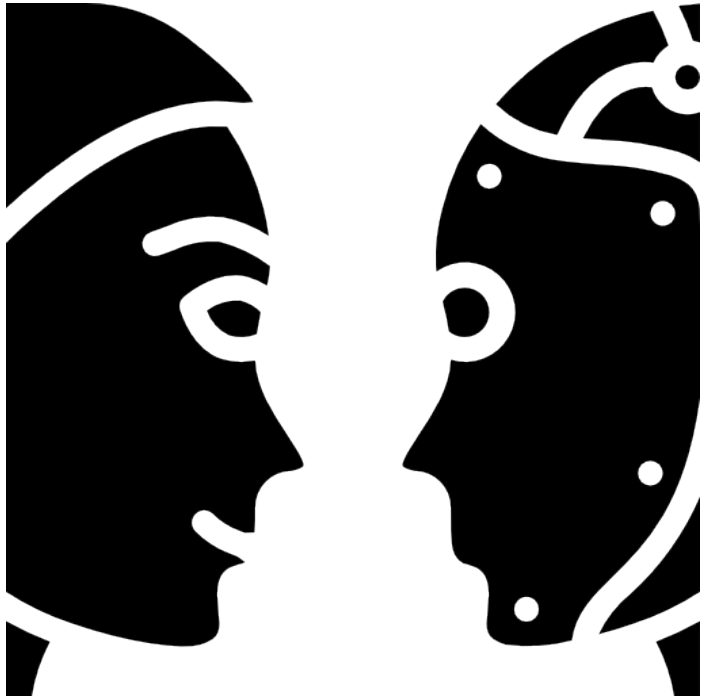
2) $B \rightarrow A : r_c$
   B sends defender A a response
   (an attempt at performing the challenge)

3) $A : V(r_c) \in \{pass, fail\}$

Finally, A verifies the challenge:

1) **Realism**: Did the model break?
2) **Identity**: Did the RT-DF get turned off?
3) **Task**: Was the task performed?
4) **Time**: Was the response performed in real-time?

…all 4 must pass



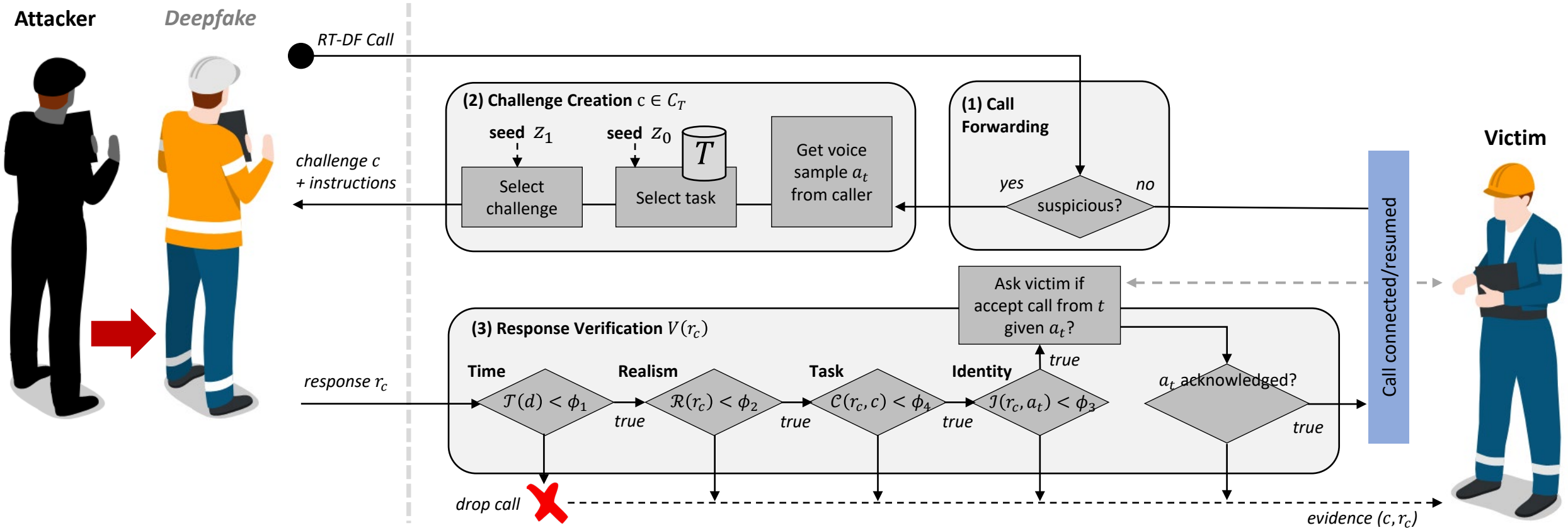
How It Works

Captcha Challenges for Voice RT-DF

Task (T)	Acronym	Usability	Hardness				Weakness	Effectiveness	
			Realism	Identity	Task	Time	Evasions	Naive Attacker	Advanced Attacker
Clear Throat	CT	●	●	○	●	●		●	○
Hold Musical Note	HN	●	○	○	●	●		●	●
Hum Tune	HT	●	●	●	●	●		●	●
Laugh	L	○	●	●	●	●		●	●
Mimic Speaking Style	MS	○	●	●	○	●		○	○
Repeat Accent	R	○	●	●	○	●		○	○
Sing	S	●	●	●	●	●		●	●
Speak with Emotion	SE	●	●	●	○	●		●	●
Yawn	Y	○	●	○	●	●		●	●
Blow Noises	BN	●	●	–	●	●	bypass	●	–
Blow on Mic	BM	○	●	–	●	●	bypass	●	–
Clap	Cl	●	○	–	●	●	bypass	●	–
Click Tongue	Clk	●	●	–	●	●	bypass	●	–
Cough	Co	●	●	–	●	●	bypass	●	–
Horse Lips	HL	○	●	–	●	●	bypass	●	–
Knock	K	○	○	–	●	●	bypass	●	–
Playback Audio	PA	–	●	–	●	●	bypass	●	–
Raspberry	R	●	●	–	●	●	bypass	●	–
Sound Effect	SFX	●	●	–	●	●	bypass	●	–
Touch Mic	TM	○	●	–	●	●	bypass	●	–
Type	T	○	●	–	●	●	bypass	●	–
Whistle	W	–	●	–	●	●	bypass	●	–
Talk & Clap	T&C	○	●	●	●	●	mix	●	–
Talk & Knock	T&K	○	●	●	●	●	mix	●	–
Talk & Playback	P	–	●	●	●	●	mix	●	–
Talk with Tones	TT	●	●	●	●	●	mix	●	●
Vary Speed	VS	●	●	●	○	●	mix	●	●
Vary Volume	V	●	●	●	○	●	mix	●	●

●: high, ○: medium, –: low

RT-DF System Design



Evaluation

Models

Realism

- *5 different DF detection models*
- *SpecRNet, OC-Resnet18, GMM-ASVspoof, PC-DARTs, LOF*
  - *Each was used as baselined too*

Task

- *GMM classifier on MFCC features*

Identity

- *Anomaly detector based on voice embeddigns taken from a pretrained voice recog. model*

Datasets

	Real: $\mathcal{D}_{real}$	**Fake:** $\mathcal{D}_{fake}$
Speech	2498	1821

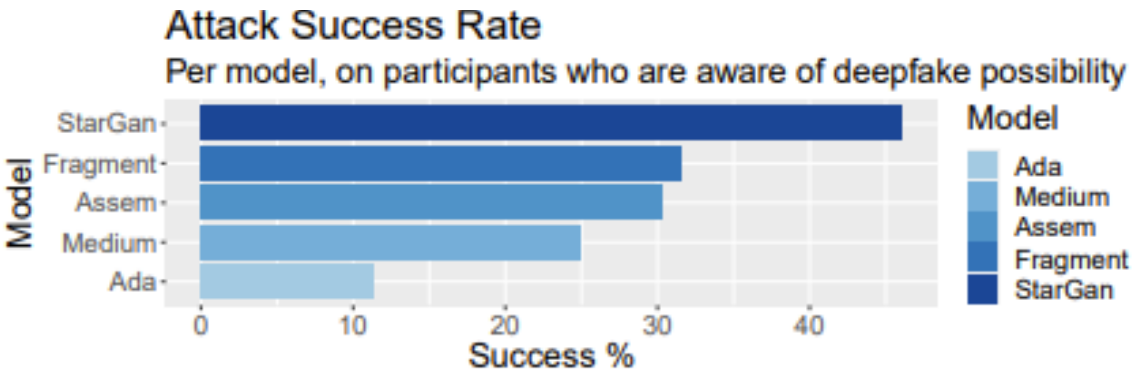
	Real: $\mathcal{D}_{real,r}$	**Fake:** $\mathcal{D}_{fake,r}$
Repeat Accent (R)	98	570
Clap (Cl)	99	551
Cough (Co)	537	3,401
Speak with Emotion (SE)	98	532
Hum Tune (HT)	593	3,325
Playback Audio (P)	601	3,420
Sing (S)	595	334
Vary Speed (VS)	98	570
Vary Volume (V)	598	3,420

	Real	**Fake**
ASVspoof-DF	22,617	15,000
RITW	19,963	11,816

Evaluation

RT-DF Attacks (2019-2021)

- *StarGANv2*
- *AdaIN-VC*
- *ASSEM-VC*
- *FragmentVC*
- *MediumVC*



Some offline Voice DF can be made real-time!

Example with StarGANv2:

Evaluation

Captcha audio vs regular audio (baseline)

AUC	Baseline	R	T&C	SE	P	VS	V	S	HT	Co
SpecRNet	0.952	0.914	0.538	0.796	0.825	0.922	0.92	0.834	0.701	0.789
One-Class	0.939	**0.952**	**0.967**	**0.941**	**0.954**	**0.958**	**0.957**	**0.948**	0.896	0.832
GMM-AsvSpoof	0.949	**0.951**	**0.978**	**0.953**	**0.97**	**0.957**	**0.949**	0.928	**0.949**	0.833
PC-DARTS	0.551	**0.568**	**0.557**	**0.611**	0.507	**0.586**	**0.579**	**0.655**	**0.675**	**0.635**
LOF	0.678	0.614	**0.93**	0.635	**0.756**	**0.771**	**0.824**	0.593	**0.681**	**0.982**

EER	Baseline	R	T&C	SE	P	VS	V	S	HT	Co
SpecRNet	0.116	0.163	0.475	0.285	0.261	0.155	0.154	0.245	0.354	0.281
One-Class	0.128	**0.123**	**0.099**	0.133	**0.118**	**0.112**	**0.104**	**0.128**	0.187	0.259
GMM-AsvSpoof	0.122	**0.1**	**0.071**	**0.099**	**0.09**	**0.092**	**0.115**	0.143	0.131	0.255
PC-DARTS	0.449	**0.418**	0.494	**0.386**	0.494	**0.43**	**0.437**	**0.366**	**0.334**	**0.415**
LOF	0.326	0.419	**0.122**	0.412	**0.262**	**0.301**	**0.26**	0.38	0.382	**0.051**

Evaluation

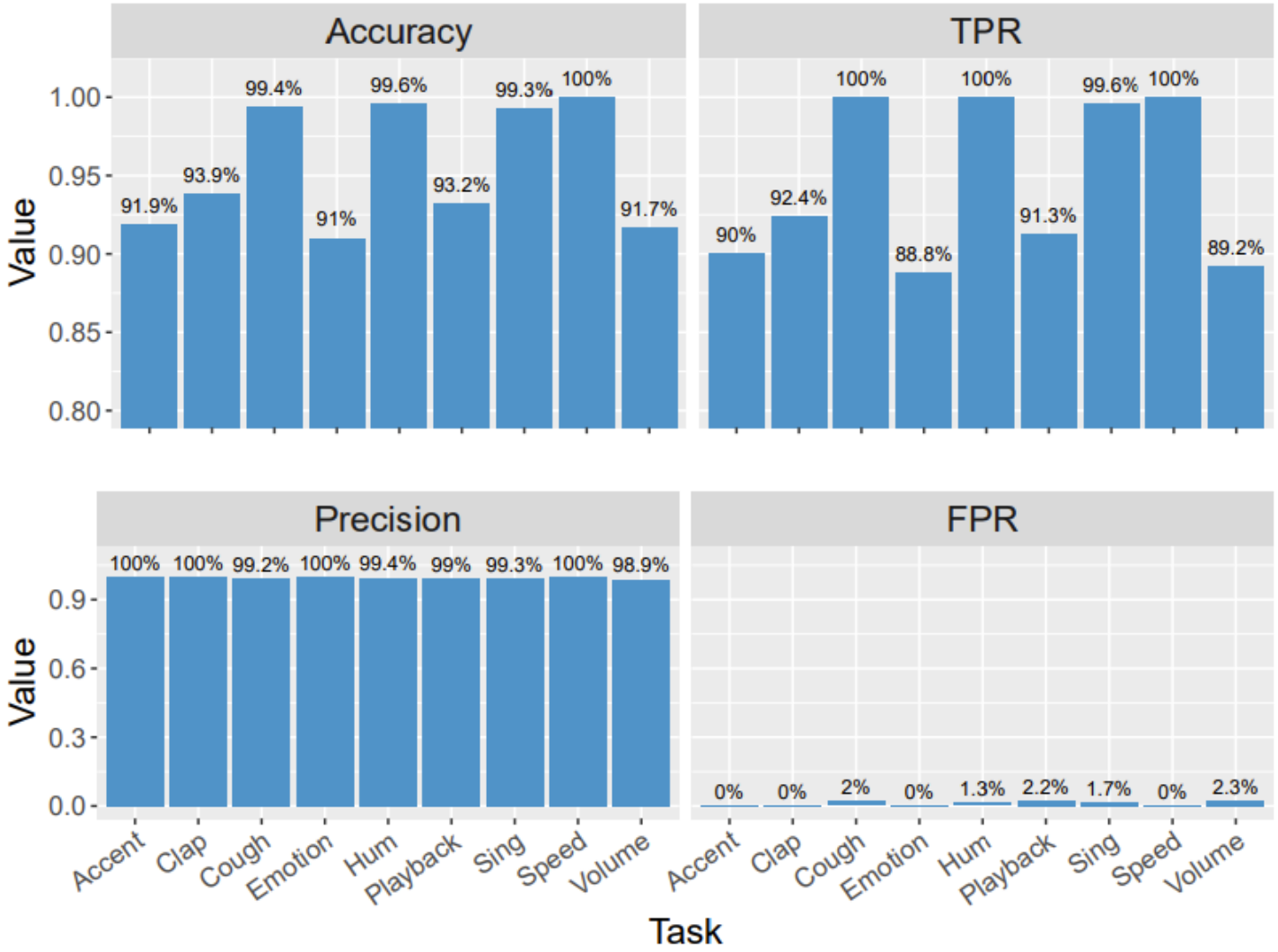
End-to-end Performance

- Random Captchas selected
- Realism, Identity and task detection models

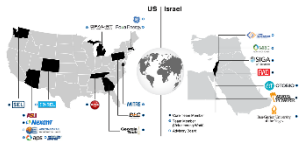
Summary:

DF-Captcha:
TPR: 0.89-1.00
FPR: 0.0-2.3
ACC: 91-100%

Baseline (best detector SpecRNet)
TPR: 0.66
ACC 71% @ FPR=0.01



Next Steps



Research

Extend system to Video RT-DFs (find captchas baselines etc…)

Publicity

- Interview on Real-time audio deepfake threat
- ASIA CCS ’23 paper (response on 22nd of March)

Commercialization

- Still no cooperation found
  - We need industry contacts
- Will publicize during the next webinar