



GRAZ – AUSTRIA
SEPTEMBER 15th – 19th 2019



ASVspoof 2019

Future Horizons in Spoofed and Fake Audio Detection

Massimiliano Todisco, EURECOM, FRANCE

Xin Wang, NII, JAPAN

Ville Vestman, University of Eastern Finland, FINLAND

Md Sahidullah, Inria, FRANCE

Héctor Delgado, EURECOM, FRANCE

Andreas Nautsch, EURECOM, FRANCE

Junichi Yamagishi, NII, JAPAN / University of Edinburgh, UK

Nicholas Evans, EURECOM, FRANCE

Tomi Kinnunen, University of Eastern Finland, FINLAND

Kong Aik Lee, NEC, JAPAN

organisers



Junichi Yamagishi
NII, Japan
Univ. of Edinburgh, UK



Massimiliano Todisco
EURECOM, France



Md Sahidullah
Inria, France



Héctor Delgado
EURECOM, France
Nuance, Spain



Xin Wang
NII, Japan



Nicholas Evans
EURECOM, France



Tomi H. Kinnunen
UEF, Finland



Kong Aik Lee
NEC, Japan



Ville Vestman
UEF, Finland



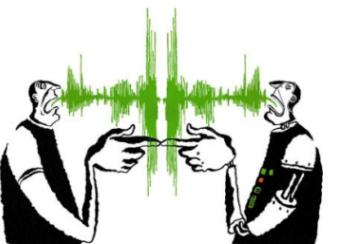
Andreas Nautsch
EURECOM, France

security issues in voice biometric

□ security in voice biometrics is becoming a necessity

The Economist Topics Print edition More Subscribe Log in or register Manage subscription

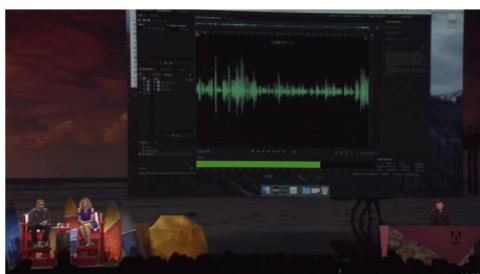
Cloning voices
Imitating people's speech patterns precisely could bring trouble
You took the words right out of my mouth



Bella Miller

BBC NEWS Home Video World UK Business Tech Science Magazine Entertainment & Arts Health More Search

Technology Adobe Voco 'Photoshop-for-voice' causes concern 7 November 2016



A new application that promises to be the "Photoshop of speech" is raising ethical and security concerns.

TECH ARTIFICIAL INTELLIGENCE Lyrebird claims it can recreate any voice using just one minute of sample audio The results aren't 100 percent convincing, but it's a sign of things to come by James Vincent @jvnc | Apr 24, 2017, 12:04pm EDT

SHARE TWEET LINKEDIN



zendesk support A beautifully simple customer service ticketing system Try it for free

Artificial intelligence is making human speech as malleable and replicable as pixels. Today, a Canadian AI startup named **Lyrebird** unveiled its first product: a set of algorithms the company claims can clone anyone's voice by listening to just a single minute of sample audio.

A few years ago this would have been impossible, but the analytic prowess of machine learning has proven to be a perfect fit for the idiosyncrasies of human speech. Using artificial intelligence, companies like Google have been able to create incredibly life-like synthesized

NOW TRENDING

The Telegraph ALL SECTIONS Technology More

This robot speech simulator can imitate anyone's voice



The machine has mimicked Barack Obama CREDIT: REX

UAB News Knowledge that will change your world Latest Updates UAB Magazine The UAB Mix UAB Reporter Media Resources

Innovation & Development

UAB research finds automated voice imitation can fool humans and machines by Katherine Shoneg



How a "voice impersonation" works

University of Alabama at Birmingham researchers have found that automated and human verification for voice-based user authentication are equally vulnerable to voice impersonation. This new research is being presented at the European Symposium on Research in Computer Security, or ESORICS, today in Vienna, Austria.

Using an off-the-shelf voice-mapping tool, the researchers developed a voice impersonation attack to successfully penetrate automated and human verification systems.

HSBC

A NEW STORAGE ARCHITECTURE FOR DATA AT SCALE THAT WON'T BREAK THE BANK

HSBC voice recognition system breached by customer's twin

BBC Click reporter Dan Simmons said his non-identical twin brother was able to fool system and gain access to account



Packt Tutorials News Learning Paths Books & Videos Podcasts Subscription

Web Development Data Mobile Programming Cloud & Networking Security Game Development IoT & Hardware

Home Data News Artificial Intelligence News Google News Initiative partners with Google AI to help 'deep fake' audio detection research

By Amrata Joshi - February 1, 2019 - 8:22 am 169 0



spoofing/presentation attacks

[ISO/IEC 30107-1:2016]

speech synthesis



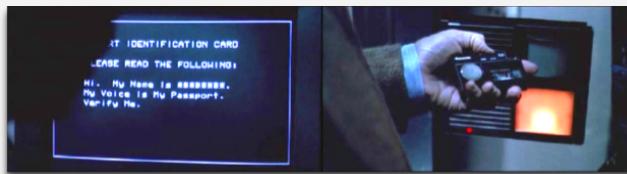
greatest threats!

voice conversion



replay

Sneakers (1992)



Universal Pictures



mobile phone



HQ loudspeaker



anechoic room
HQ loudspeaker

impersonation

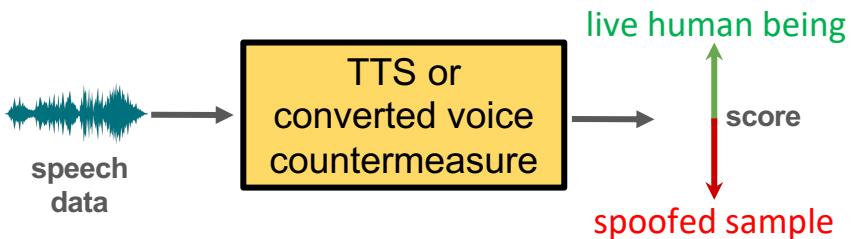
Nick here...
verify my voice!



mimicry by a human being

past ASVspoof challenge tasks

ASVspoof 2015



16 organizations participated

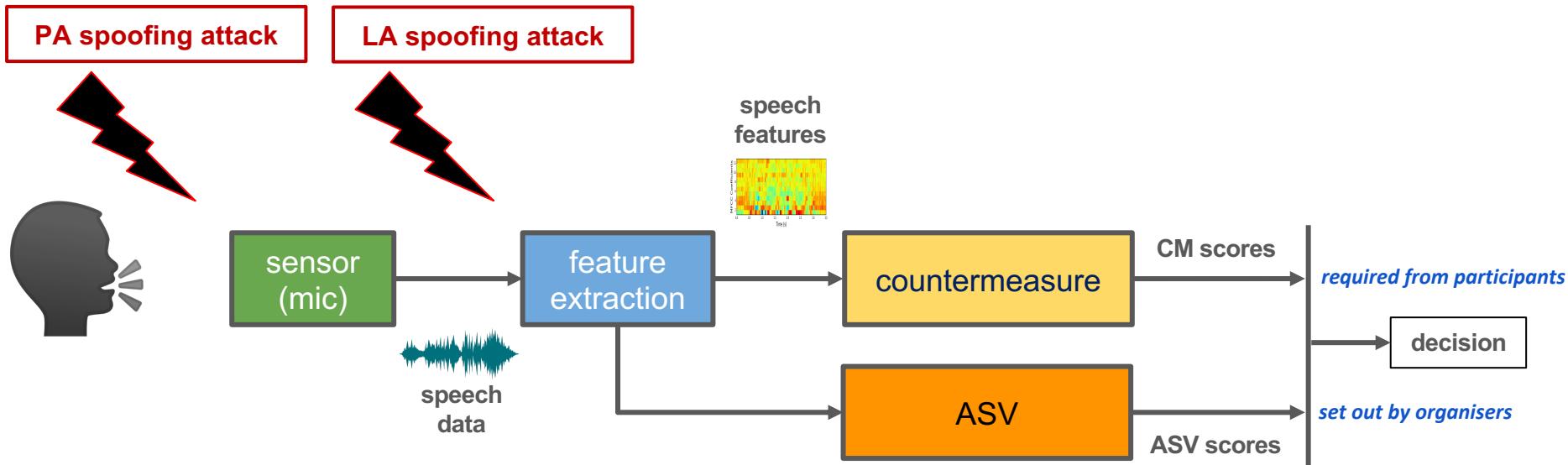
ASVspoof 2017



49 organizations participated

ASVspoof 2019 challenge

- ❑ ASV-centric
- ❑ logical access (LA)
- ❑ physical access (PA) { separately evaluated

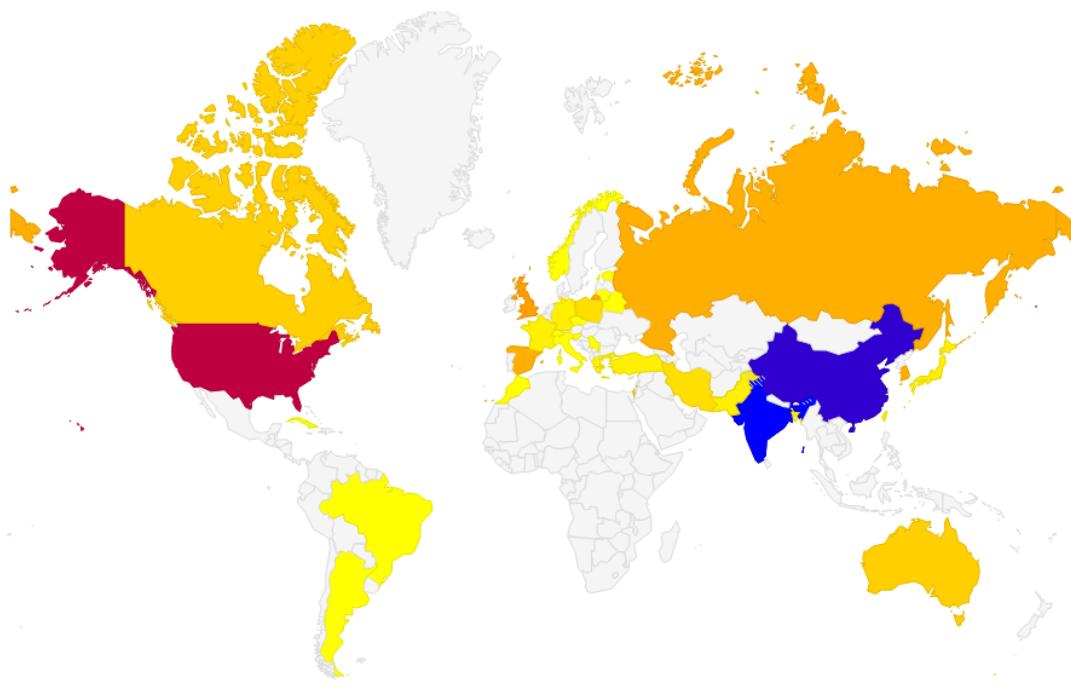
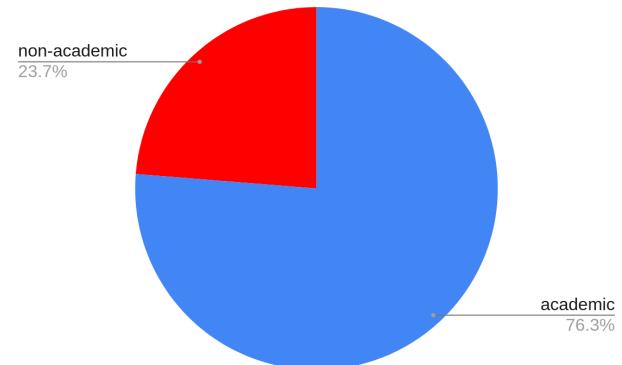


- ❑ primary metric: minimum normalized tandem Detection Cost Function (t-DCF) [1]
- ❑ secondary metric: equal error rate (EER) of *spoof - nonspoof* discrimination

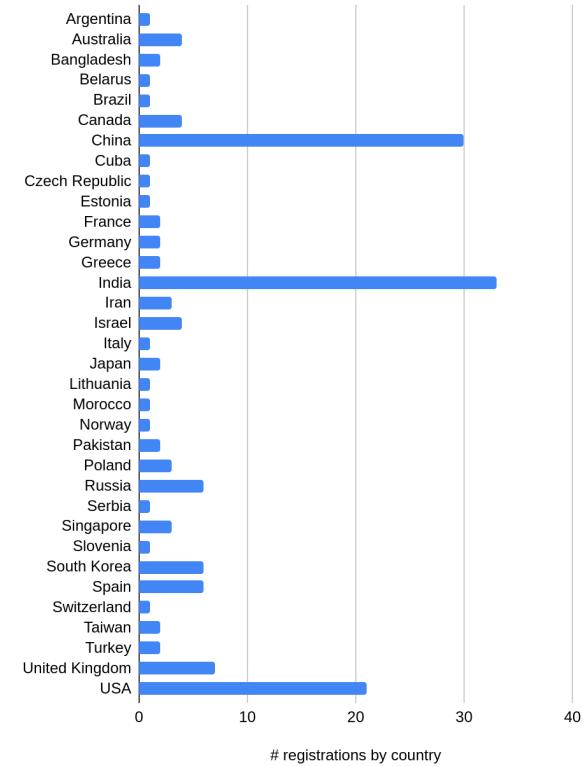
[1] T. Kinnunen, K. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, “t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification,” in Proc. Odyssey, Les Sables d’Olonne, France, June 2018.

participant statistics

- registration: 154 teams or individuals
- submitted results
 - 48 (31%) for LA scenario
 - 50 (32%) for PA scenario



1 33



database & protocols

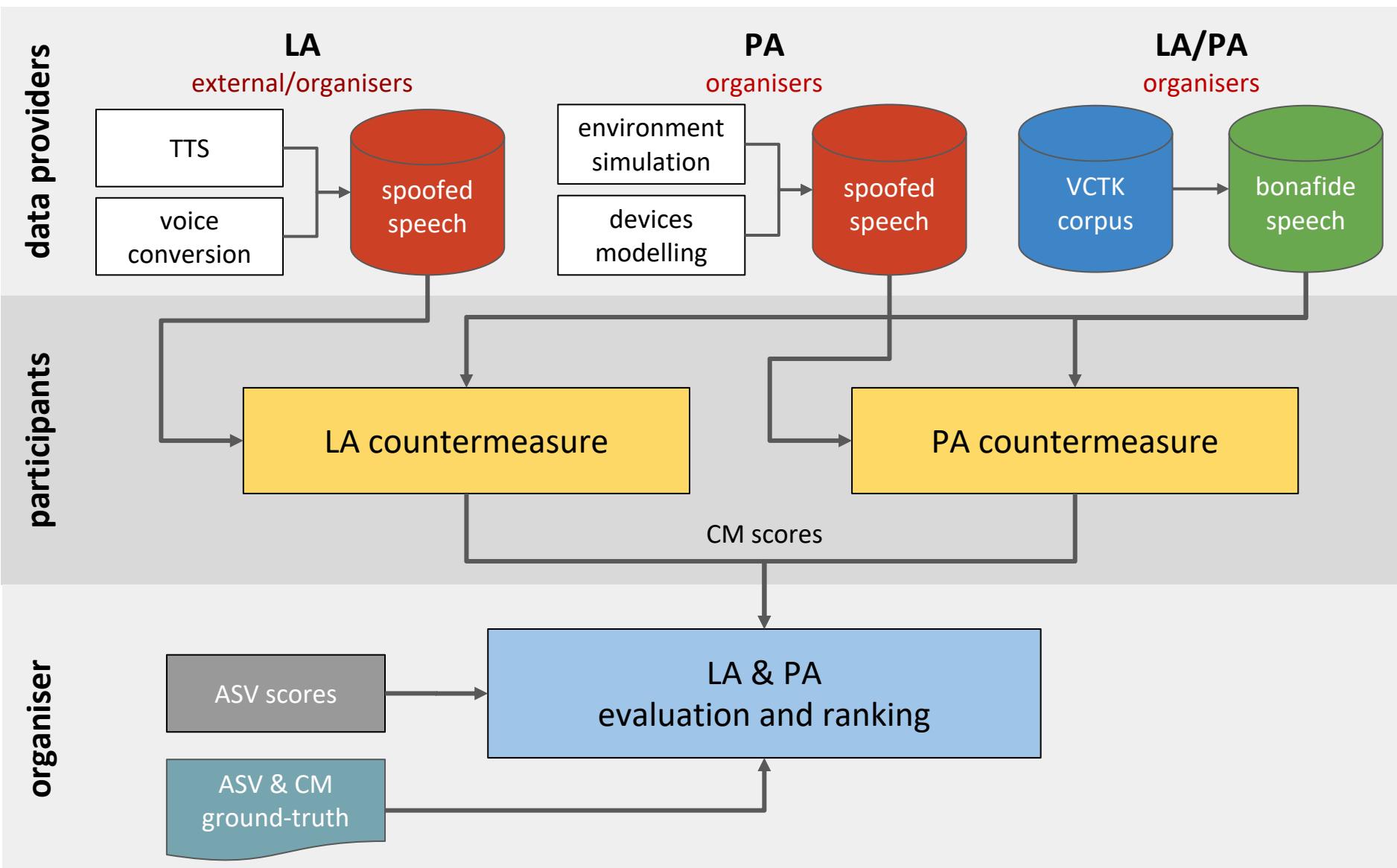
- ❑ based on **VCTK corpus** [1]
 - ❑ hemi-anechoic chamber of the University of Edinburgh
 - ❑ omni-directional head-mounted microphone (DPA 4035)
 - ❑ 96 kHz sampling frequency @ 24 bits
 - ❑ downsampled to 16 kHz @ 16 bits
- ❑ common partitions for LA and PA
 - ❑ 107 English speakers
 - ❑ speakers for eval, dev and training set
 - ❑ ASV enrolment (spks & utts)



VCTK corpus

[1] C. Veaux, J. Yamagishi, K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.

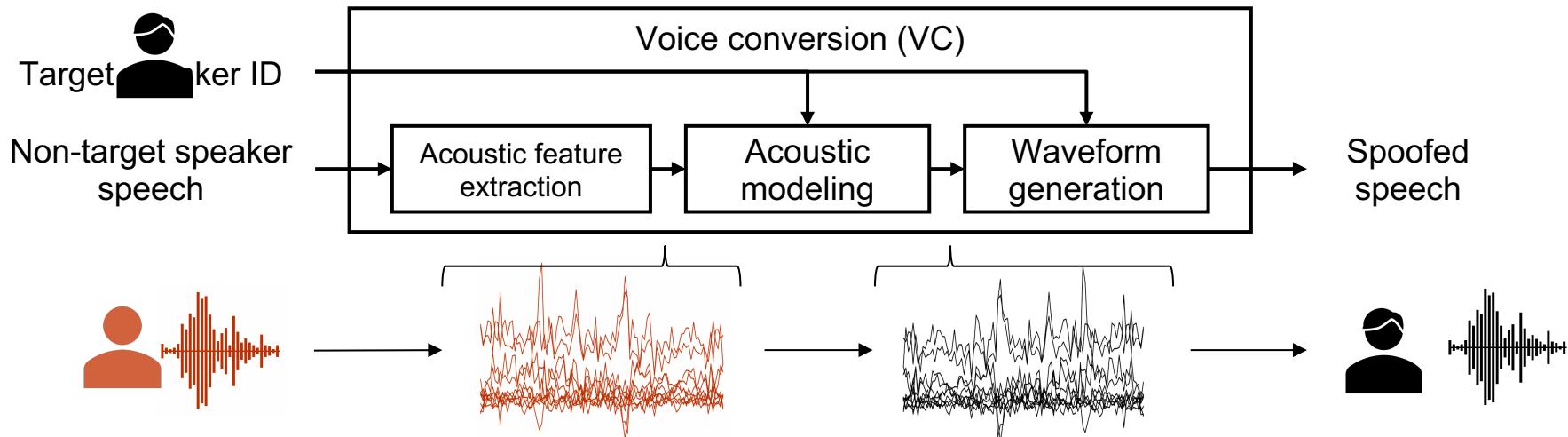
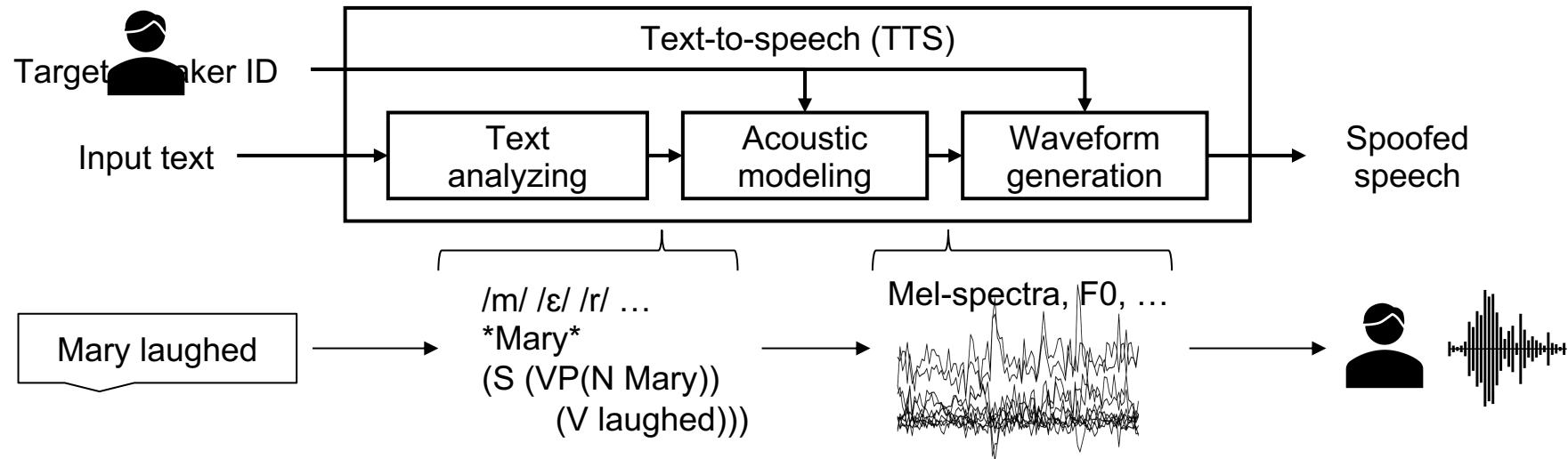
flow of the ASVspoof 2019 challenge



logical access

logical access

□ Spoofing TTS/VC systems in general



logical access

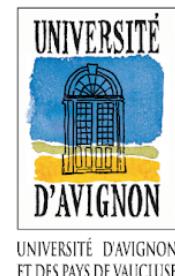
□ Spoofing TTS/VC systems in general



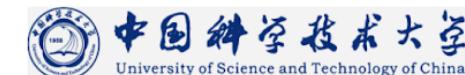
Google AI



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin



THE UNIVERSITY
of EDINBURGH



logical access

❑ Spoofing TTS/VC systems

	Category	Acoustic model	Waveform generator	
A01	TTS	VAE + AR LSTM-RNN	WaveNet	
A02	TTS	VAE + AR LSTM-RNN	WORLD	
A03	TTS	Feedforward NN	WORLD	
A04	TTS	-	Waveform concat.	←
A05	VC	VAE	WORLD	
A06	VC	GMM-UBM	Spectral filtering	←
A07	TTS	LSTM-RNN	WORLD + GAN	
A08	TTS	AR LSTM-RNN	Neural source-filter model	
A09	TTS	LSTM-RNN	Vocaine	
A10	TTS	Attention seq2seq model	WaveRNN	
A11	TTS	Attention seq2seq model	Griffin-Lim	
A12	TTS	-	WaveNet	
A13	TTS-VC	Moment matching NN	Waveform filtering	
A14	TTS-VC	LSTM-RNN	STRAIGHT	
A15	TTS-VC	LSTM-RNN	WaveNet	
A16	TTS	-	Waveform concat.	←
A17	VC	VAE	Waveform filtering	
A18	VC	i-vector/PLDA	MFCC-to-waveform	
A19	VC	GMM-UBM	Spectral filtering	←

Train & dev

Evaluation

The same TTS/VC
algorithms

logical access

❑ Spoofing TTS/VC systems

	Category	Acoustic model	Waveform generator		
A01	TTS	VAE + AR LSTM-RNN	WaveNet		Train & dev
A02	TTS	VAE + AR LSTM-RNN	WORLD		
A03	TTS	Feedforward NN	WORLD		
A04	TTS	-	Waveform concat.		
A05	VC	VAE	WORLD		
A06	VC	GMM-UBM	Spectral filtering		
A07	TTS	LSTM-RNN	WORLD + GAN		Evaluation
A08	TTS	AR LSTM-RNN	Neural source-filter model		
A09	TTS	LSTM-RNN	Vocaine		
A10	TTS	Attention seq2seq model	WaveRNN		
A11	TTS	Attention seq2seq model	Griffin-Lim		
A12	TTS	-	WaveNet		
A13	TTS-VC	Moment matching NN	Waveform filtering		
A14	TTS-VC	LSTM-RNN	STRAIGHT		
A15	TTS-VC	LSTM-RNN	WaveNet		
A16	TTS	-	Waveform concat.		Varied or improved TTS/VC algorithms
A17	VC	VAE	Waveform filtering		
A18	VC	i-vector/PLDA	MFCC-to-waveform		
A19	VC	GMM-UBM	Spectral filtering		

logical access

❑ Spoofing TTS/VC systems

	Category	Acoustic model	Waveform generator	
A01	TTS	VAE + AR LSTM-RNN	WaveNet	Train & dev
A02	TTS	VAE + AR LSTM-RNN	WORLD	
A03	TTS	Feedforward NN	WORLD	
A04	TTS	-	Waveform concat.	
A05	VC	VAE	WORLD	
A06	VC	GMM-UBM	Spectral filtering	
A07	TTS	LSTM-RNN	WORLD + GAN	Evaluation
A08	TTS	AR LSTM-RNN	Neural source-filter model	
A09	TTS	LSTM-RNN	Vocaine	
A10	TTS	Attention seq2seq model	WaveRNN	
A11	TTS	Attention seq2seq model	Griffin-Lim	
A12	TTS	-	WaveNet	
A13	TTS-VC	Moment matching NN	Waveform filtering	
A14	TTS-VC	LSTM-RNN	STRAIGHT	
A15	TTS-VC	LSTM-RNN	WaveNet	Unknown TTS/VC algorithms
A16	TTS	-	Waveform concat.	
A17	VC	VAE	Waveform filtering	
A18	VC	i-vector/PLDA	MFCC-to-waveform	
A19	VC	GMM-UBM	Spectral filtering	

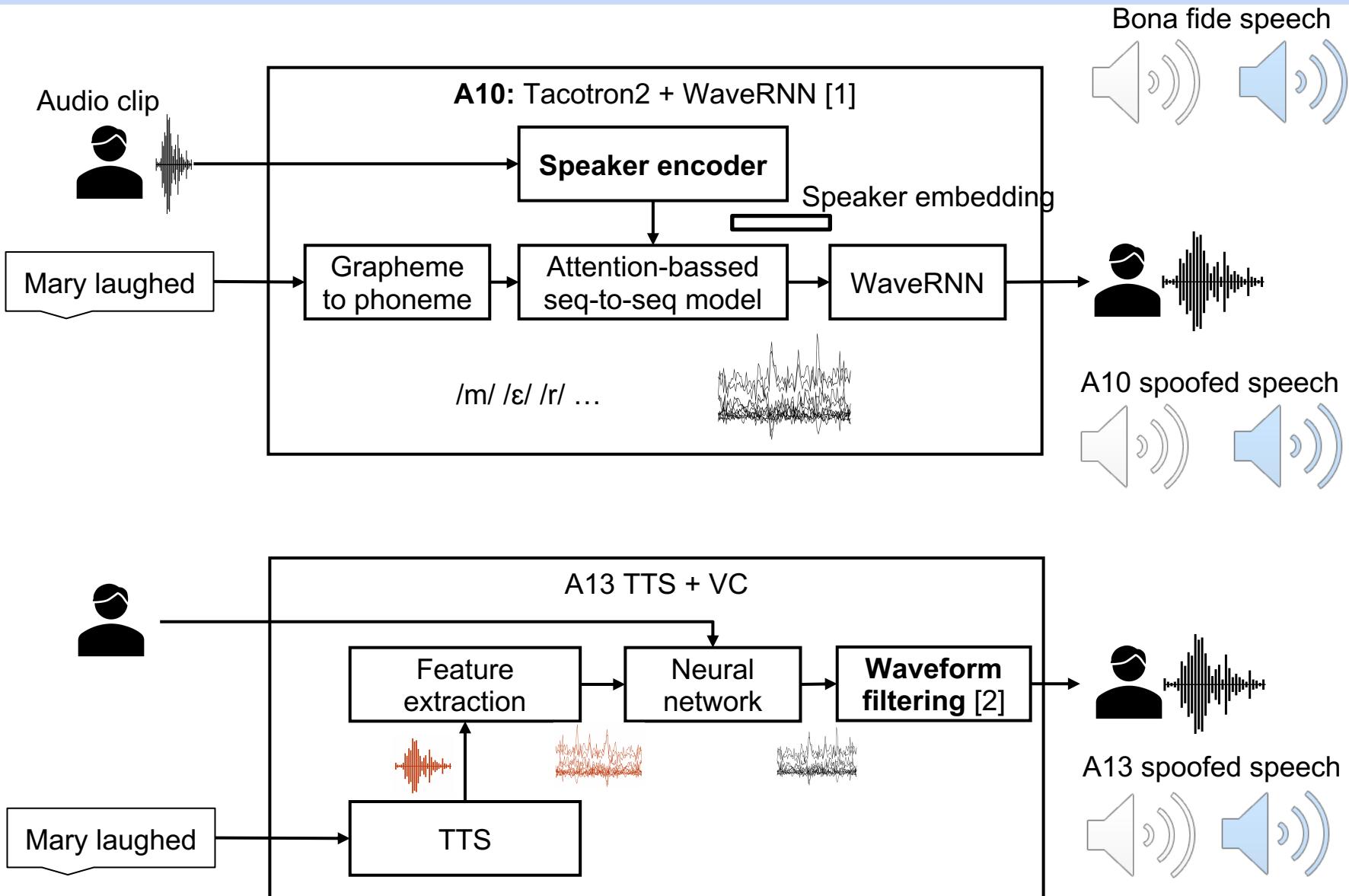
logical access

❑ Spoofing TTS/VC systems

❑ Without spoofed speech: ASV-EER = 2.48%

	Category	Acoustic model	Waveform generator	ASV-EER	CM-EER
Dev.	A01	TTS	VAE + AR LSTM-RNN	WaveNet	24.52
	A02	TTS	VAE + AR LSTM-RNN	WORLD	15.04
	A03	TTS	Feedforward NN	WORLD	56.94
	A04	TTS	-	Waveform concat.	63.02
	A05	VC	VAE	WORLD	21.90
	A06	VC	GMM-UBM	Spectral filtering	10.11
Eva.	A07	TTS	LSTM-RNN	WORLD + GAN	59.68
	A08	TTS	AR LSTM-RNN	Neural source-filter model	40.39
	A09	TTS	LSTM-RNN	Vocaine	8.38
	A10	TTS	Attention seq2seq model	WaveRNN	57.73
	A11	TTS	Attention seq2seq model	Griffin-Lim	59.64
	A12	TTS	-	WaveNet	46.18
	A13	TTS-VC	Moment matching NN	Waveform filtering	46.78
	A14	TTS-VC	LSTM-RNN	STRAIGHT	64.01
	A15	TTS-VC	LSTM-RNN	WaveNet	58.85
	A16	TTS	-	Waveform concat.	64.52
	A17	VC	VAE	Waveform filtering	3.92
	A18	VC	i-vector/PLDA	MFCC-to-waveform	7.35
	A19	VC	GMM-UBM	Spectral filtering	14.58
					0.06

logical access



[1] Jia, Y..et. al, Transfer learning from speaker verification to multispeaker text-to-speech synthesis, in: NIPS, pp. 4480–4490.

[2] Kobayashi, et al, Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential, 2018

physical access

physical access

- ❑ based upon *simulated* and carefully controlled acoustic and replay configurations
- ❑ room acoustics simulation under varying source/receiver positions using image-source method for room impulse response [1,2]
- ❑ devices modelling using the generalised polynomial Hammerstein model and the Synchronized Swept Sine tool [3]

[1] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[2] E. Vincent. (2008) Roomsimove. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

[3] A. Novak, P. Lotton, and L. Simon, "Synchronized swept-sine: Theory, application, and implementation," *J. Audio Eng. Soc.*, vol. 63, no. 10, pp. 786–798, 2015 .

physical access - replay attack definition

❑ environment

- ❑ room size**
- ❑ convolutive noise**
- ❑ recording distance between bonafide user and ASV**
- ❑ additive noise**

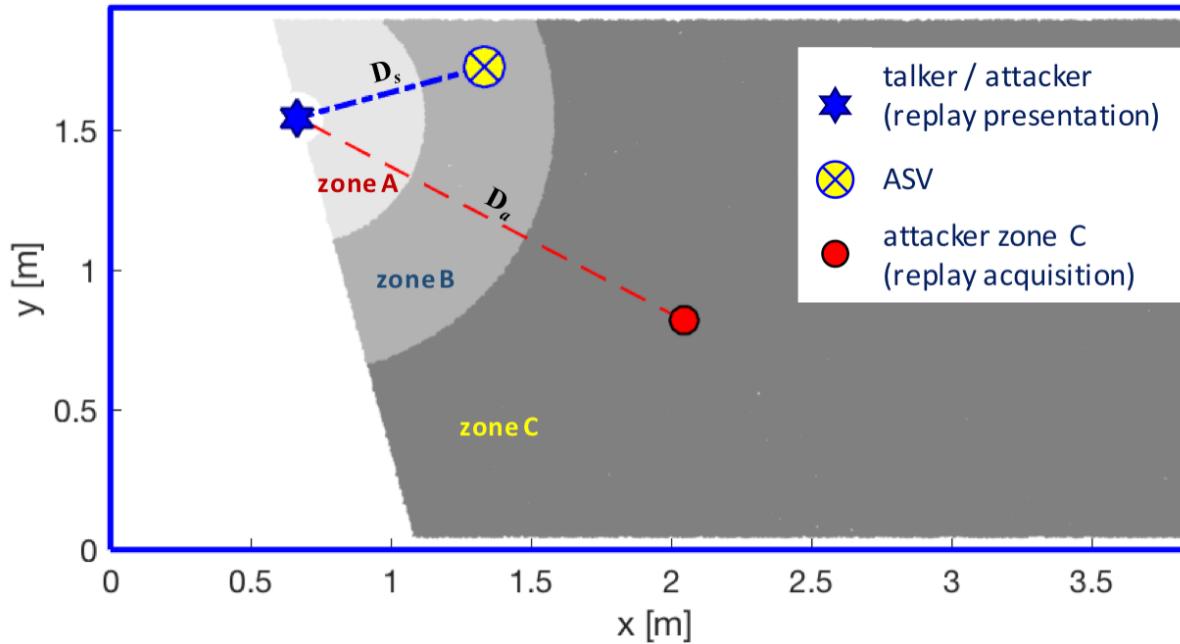
❑ replay acquisition

- ❑ recording distance between bonafide user and attacker**
- ❑ device quality (microphone)**

❑ replay presentation

- ❑ device quality (loudspeaker)**
- ❑ playback distance between attacker and ASV**

physical access - environment & attack definition



environment definition

- defined as a triplet (S, R, D_s)
- the set $(a, b, c) \rightarrow$ categorical value

attack definition

- defined as duple (D_a, Q)
- the set $(A, B, C) \rightarrow$ categorical value

device quality (Q)

- occupied bandwidth (OB) [kHz]
- lower bound of OB (minF) [Hz]
- linear/nonlinear OB power difference (linearity) [dB]

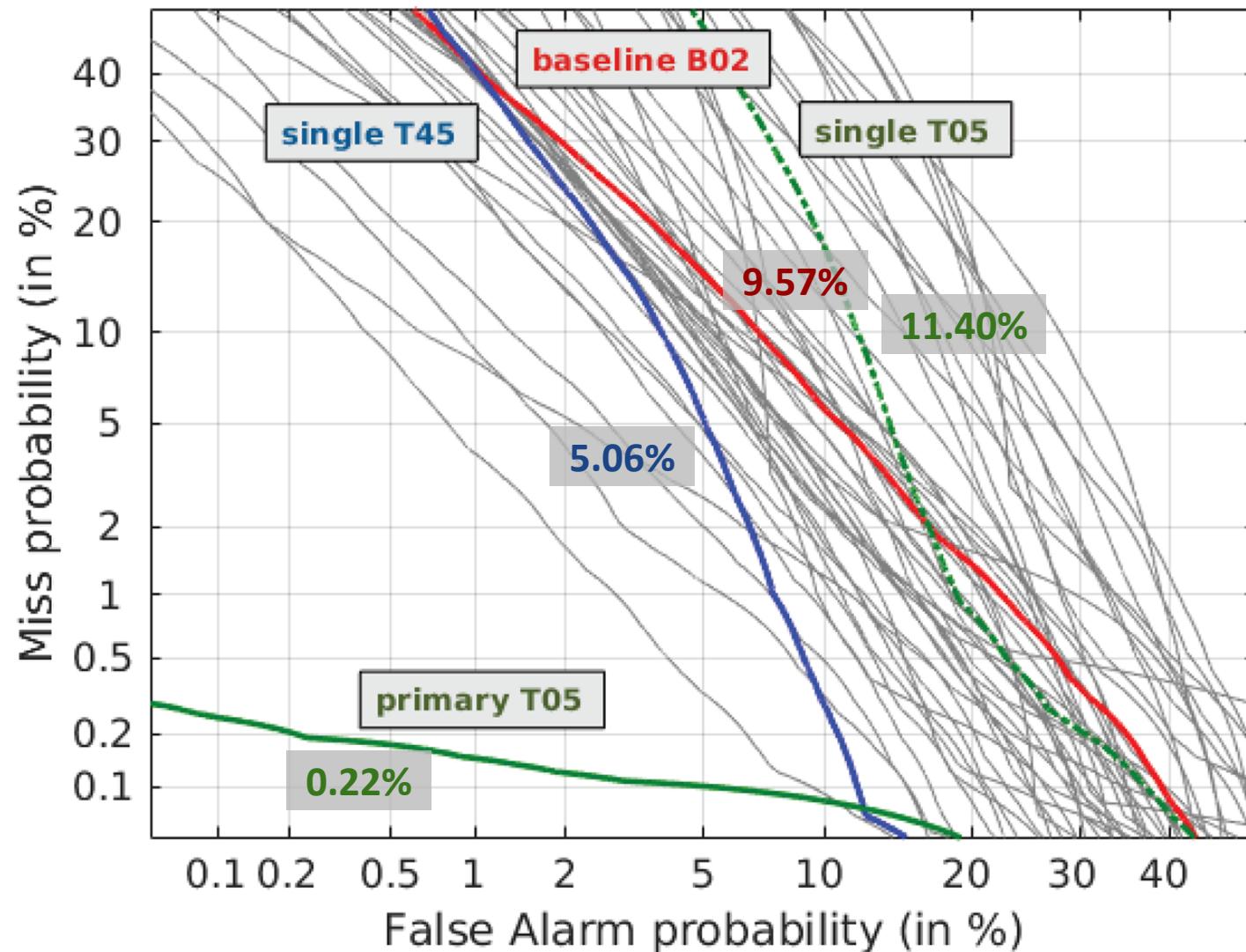
Environment definition	labels		
	a	b	c
S: Room size (square meters)	2-5	5-10	10-20
R: T60 (ms)	50-200	200-600	600-1000
D_s: Talker-to-ASV distance (cm)	10-50	50-100	100-150

Attack definition	labels		
	A	B	C
D_a: Attacker-to-talker distance (cm)	10-50	50-100	> 100
Q: Replay device quality	perfect	high	low

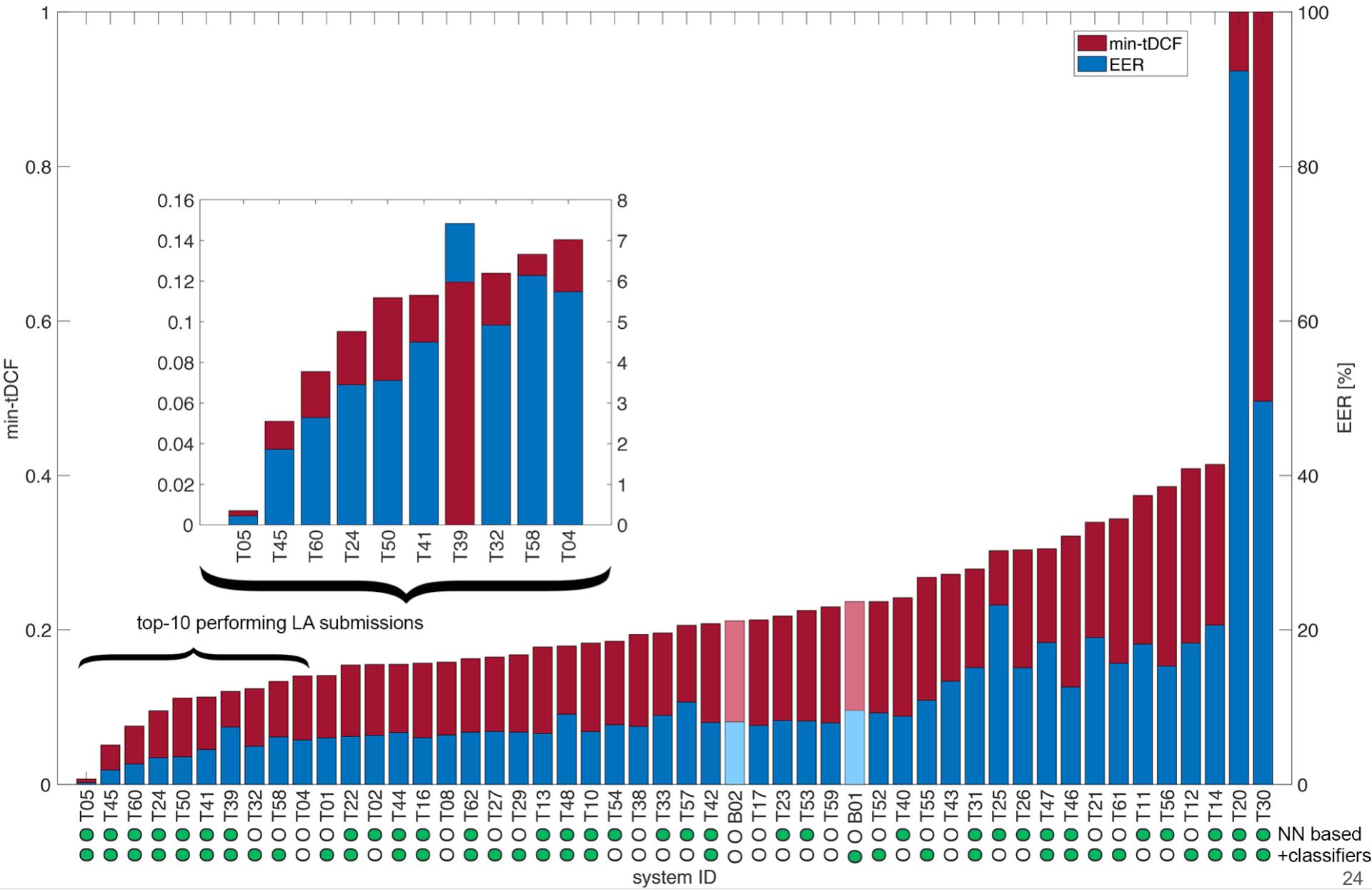
Replay device quality	OB (kHz)	minF (Hz)	linearity (dB)
Perfect	inf	0	inf
High	> 10	< 600	> 100
Low	< 10	> 600	< 100

LA challenge results

LA - common primary submissions' results



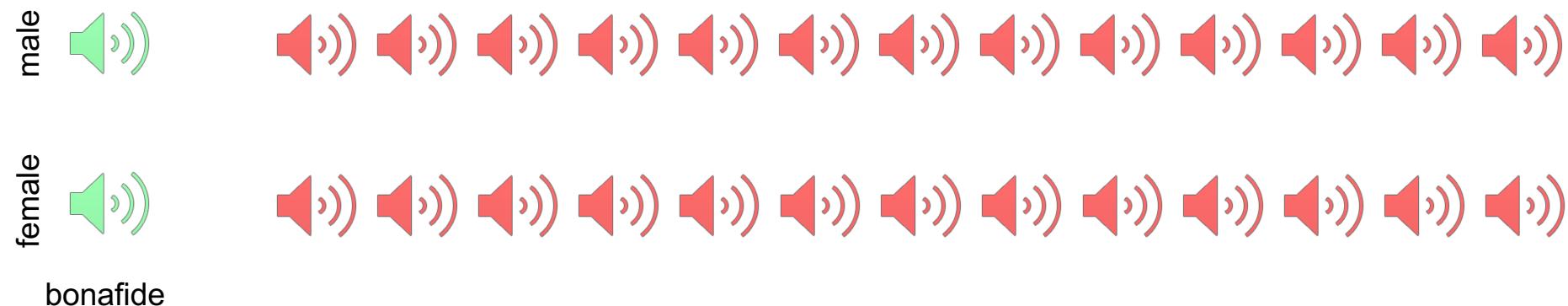
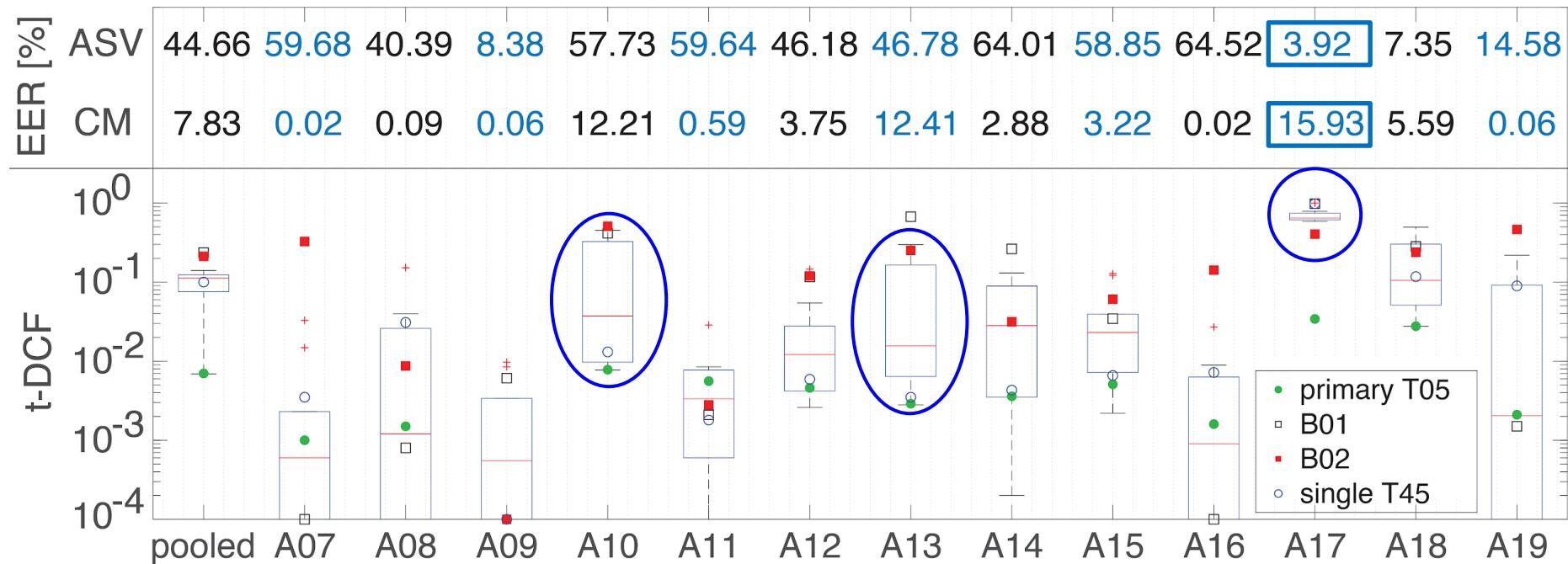
LA - common primary submissions' results



LA - 13 attacks breakdown

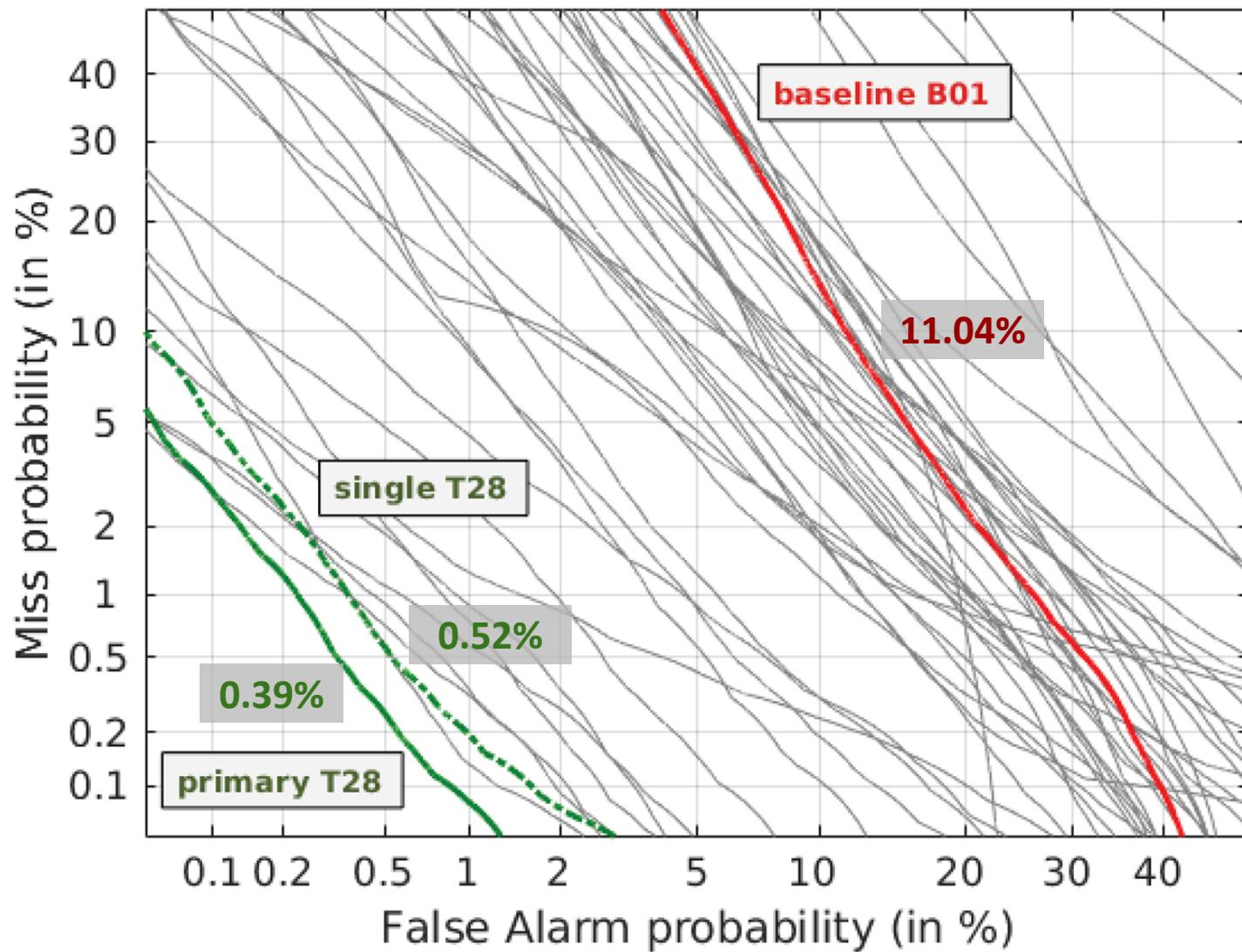
ASV only zero-effort impostors → EER = 2.48%

CM → median over all the primary submissions
t-DCF → top-10 primary submission

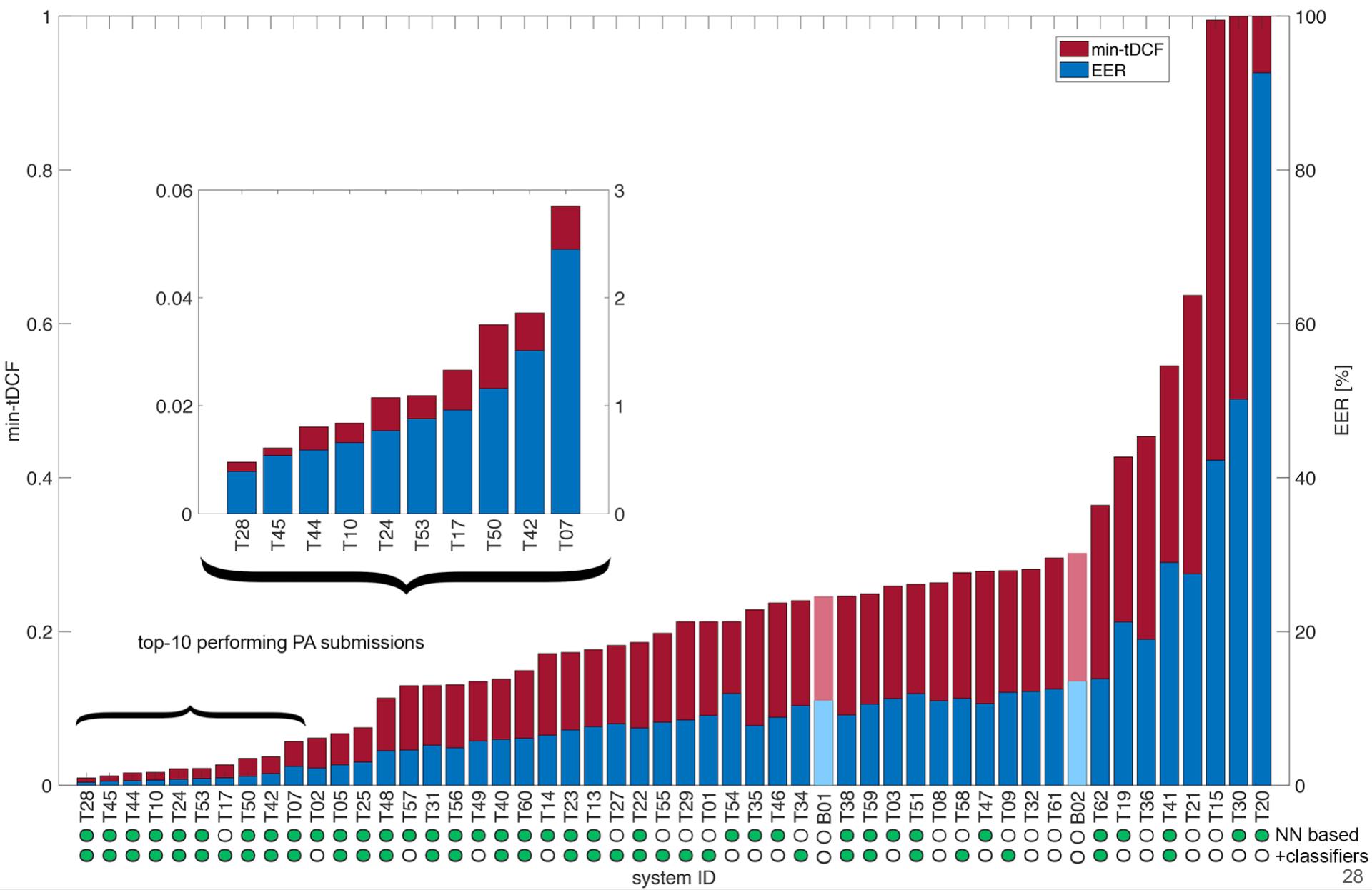


PA challenge results

PA - common primary submissions' results



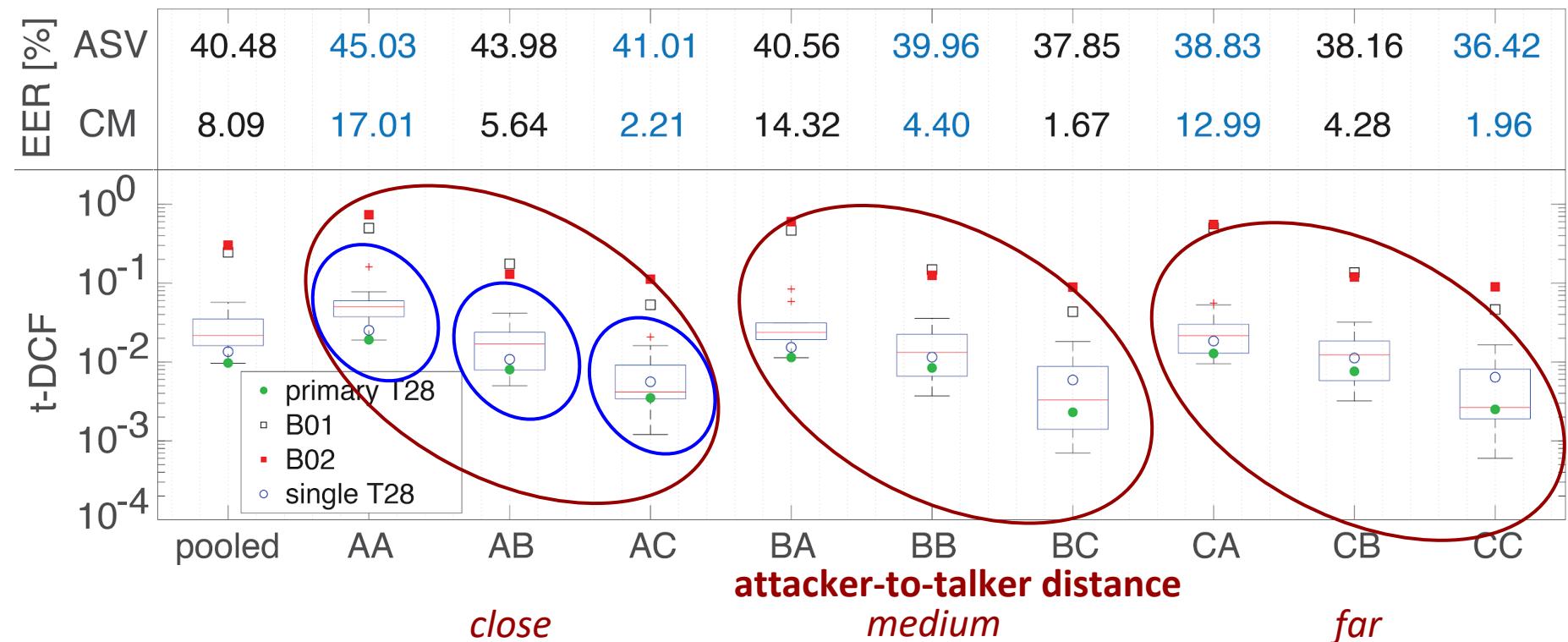
PA - common primary submissions' results



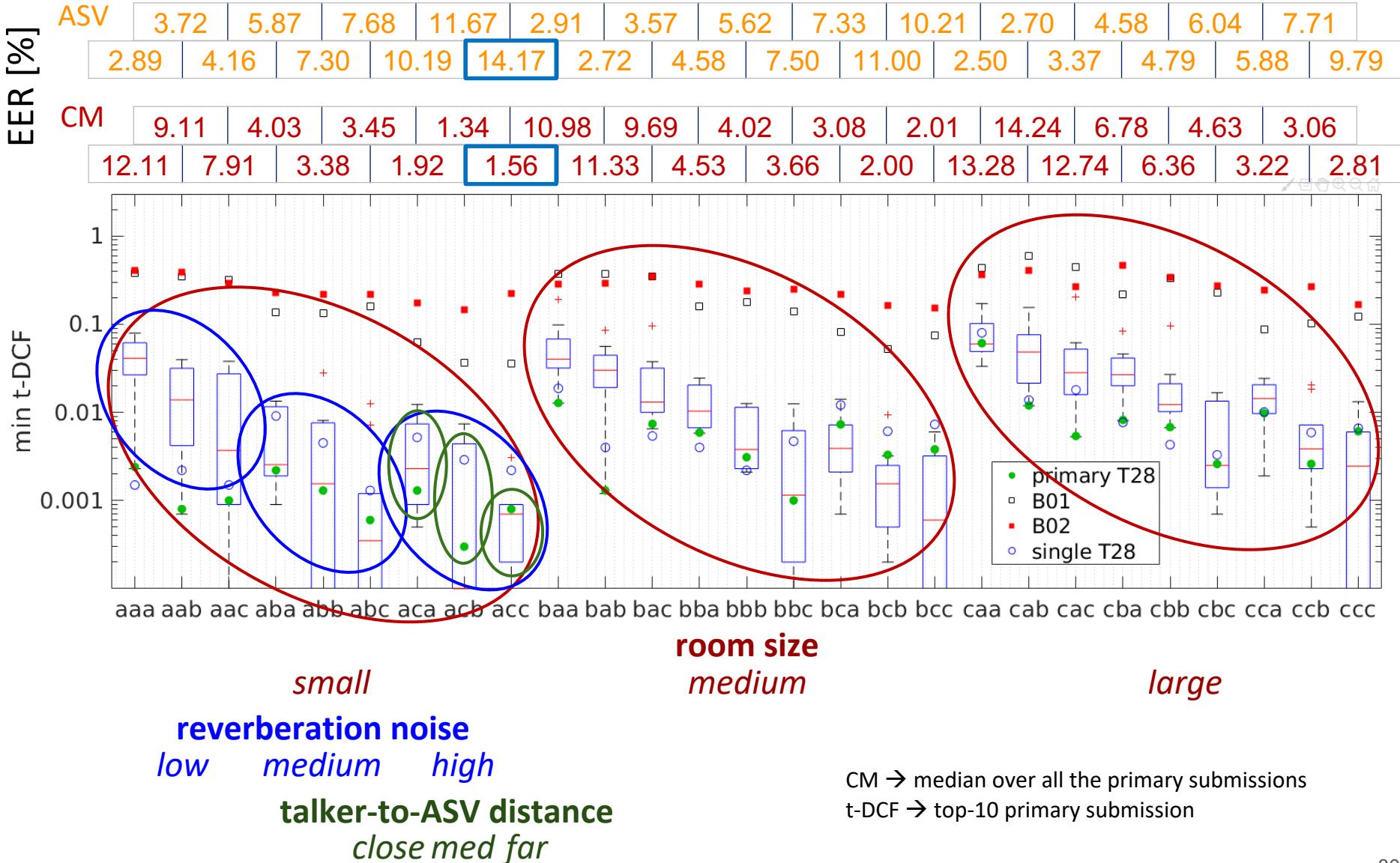
PA - 9 attacks breakdown

ASV only zero-effort impostors → EER = 6.47%

CM → median over all the primary submissions
t-DCF → top-10 primary submission



PA - 27 environments breakdown



real PA
...hidden tracks in the album

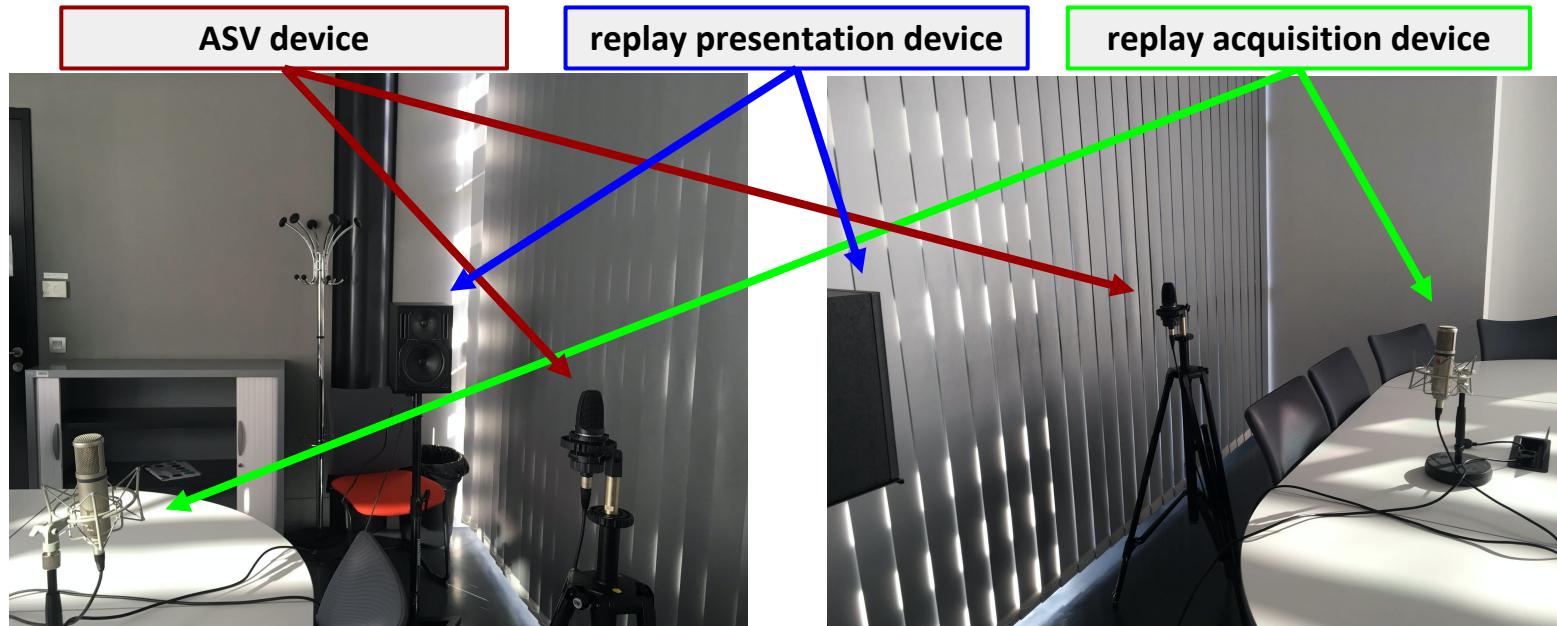
real PA - laboratory replay attacks

- ❑ small set of audio files recorded and replayed in 3 different labs
- ❑ a total of 2700 bonafide and spoof utterances
- ❑ ASVspoof 2019 real PA database contains additive noise
- ❑ 48 kHz sampling frequency @ 24 bits
- ❑ downsampled to 16 kHz @ 24 bits



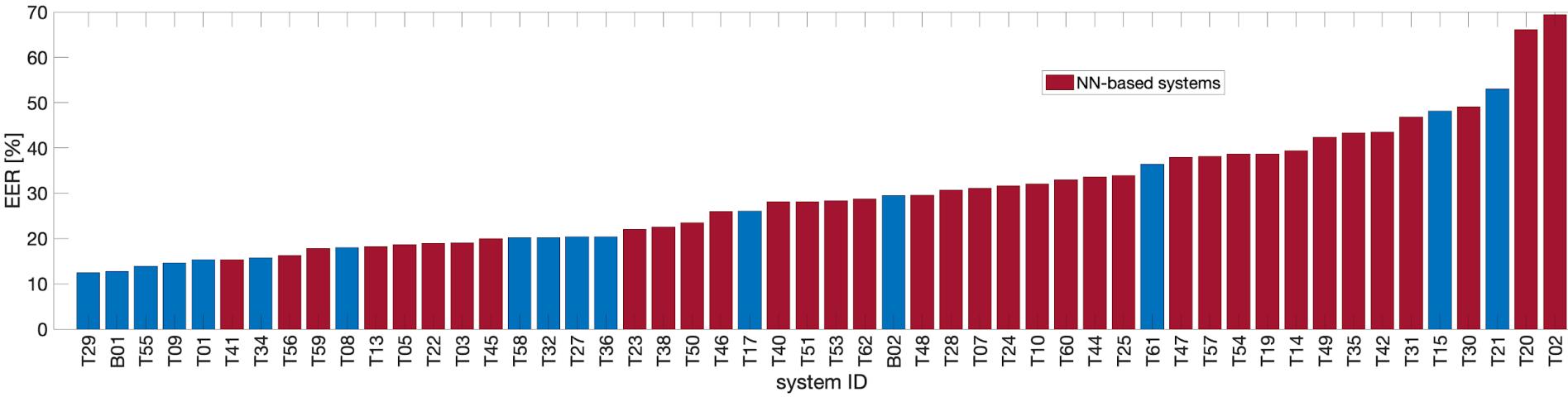
real PA

lab	#spk	#utt	#asv	#acq. dev	#pres. dev	#bonafide trials	#spoof trials
l1	6	15	1	1	4	90	360
l2	10	15	2	1	2	300	600
l3	10	15	1	8	1	150	1200



real PA - common primary submissions' results

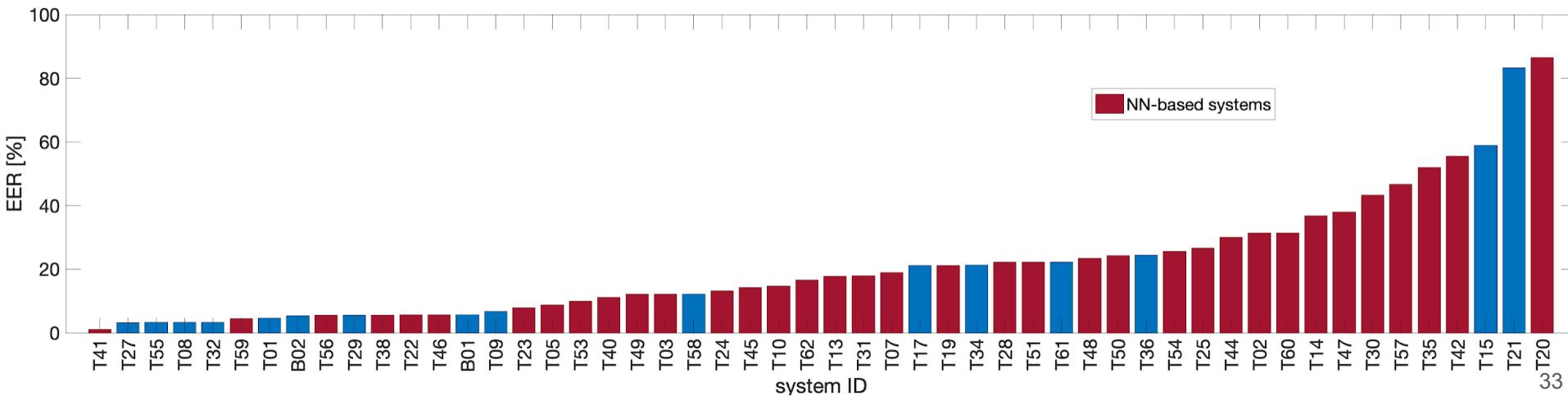
EER pooled over all attacks and labs



single quietest lab (I1)

EER pooled over attacks

Bluetooth speaker, desk loudspeaker and mobile phone



conclusion

- ❑ ASV-centric assessment
- ❑ LA and PA scenarios
- ❑ impressive results
- ❑ DNN and ensemble of classifiers, but...
 - ❑ some overfitting / lack of generalisation
 - ❑ latest TTS and VC techniques
 - ❑ real replay
- ❑ industrial and academic participation

thank you for your attention