GRAZ – AUSTRIA
SEPTEMBER 15th – 19th 2019

INTERSPEECH 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

EURECOM Sophia Antipolis
NII National Institute of Informatics
UNIVERSITY OF EASTERN FINLAND
Inria
THE UNIVERSITY of EDINBURGH
NEC

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

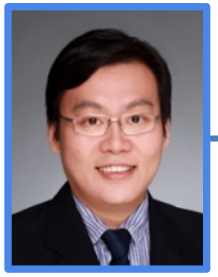
NII, Japan

Nicholas Evans

EURECOM, France

Tomi H. Kinnunen

UEF, Finland



Kong Aik Lee

NEC, Japan

Ville Vestman

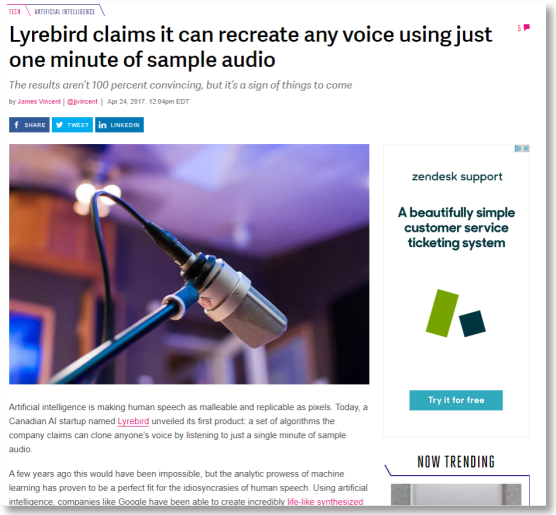
UEF, Finland

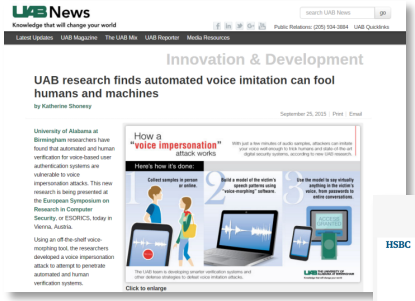
Andreas Nautsch

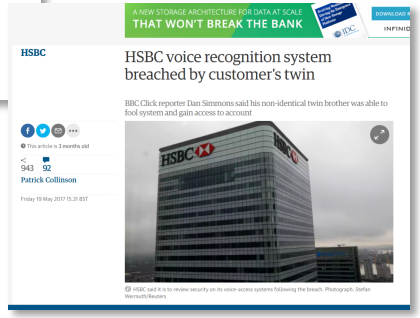
EURECOM, France

security issues in voice biometric

- security in voice biometrics is becoming a necessity





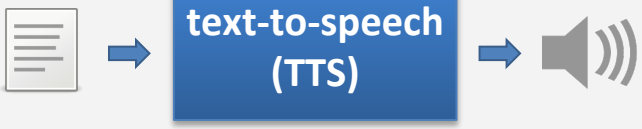


spoofing/presentation attacks

[ISO/IEC 30107-1:2016]

greatest threats!

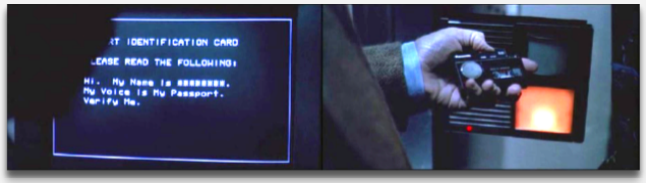
speech synthesis



voice conversion

replay

Sneakers (1992) Universal Pictures



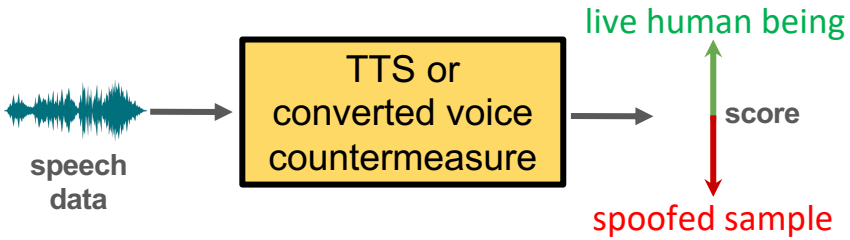
office	meeting room	anechoic room
		
mobile phone	HQ loudspeaker	HQ loudspeaker

impersonation

mimicry by a human being

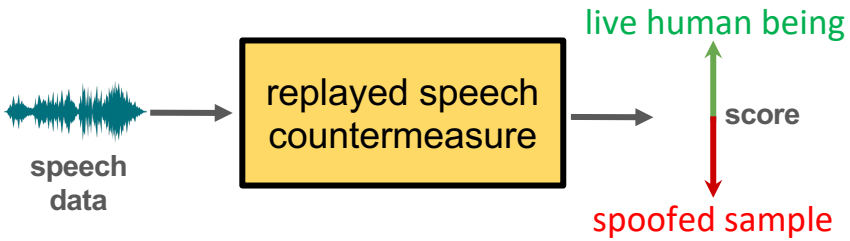
past ASVspoof challenge tasks

ASVspoof 2015



16 organizations participated

ASVspoof 2017

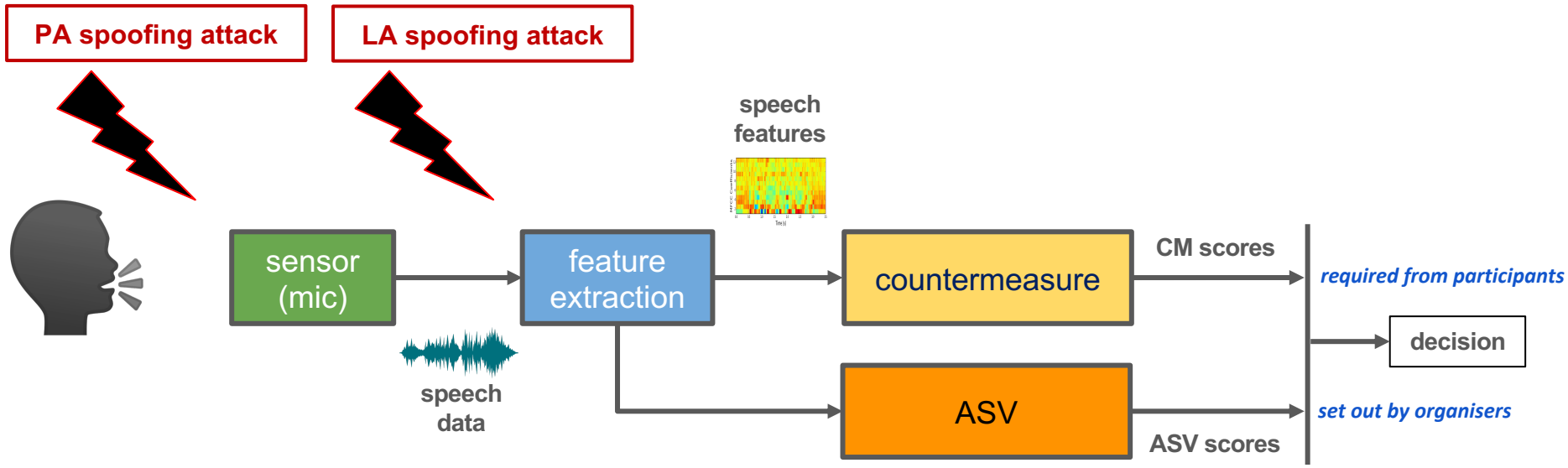


49 organizations participated

ASVspoof 2019 challenge

- **ASV-centric**
- logical access (LA)
- physical access (PA)

{ separately evaluated

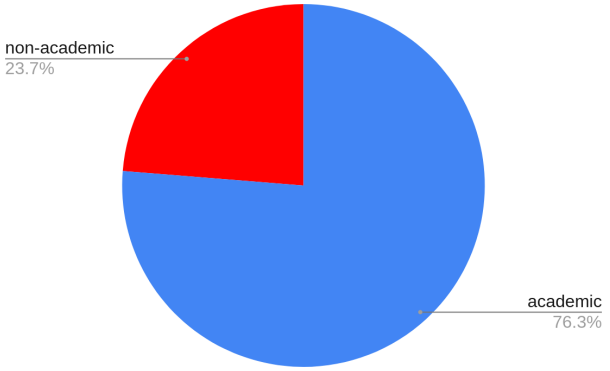


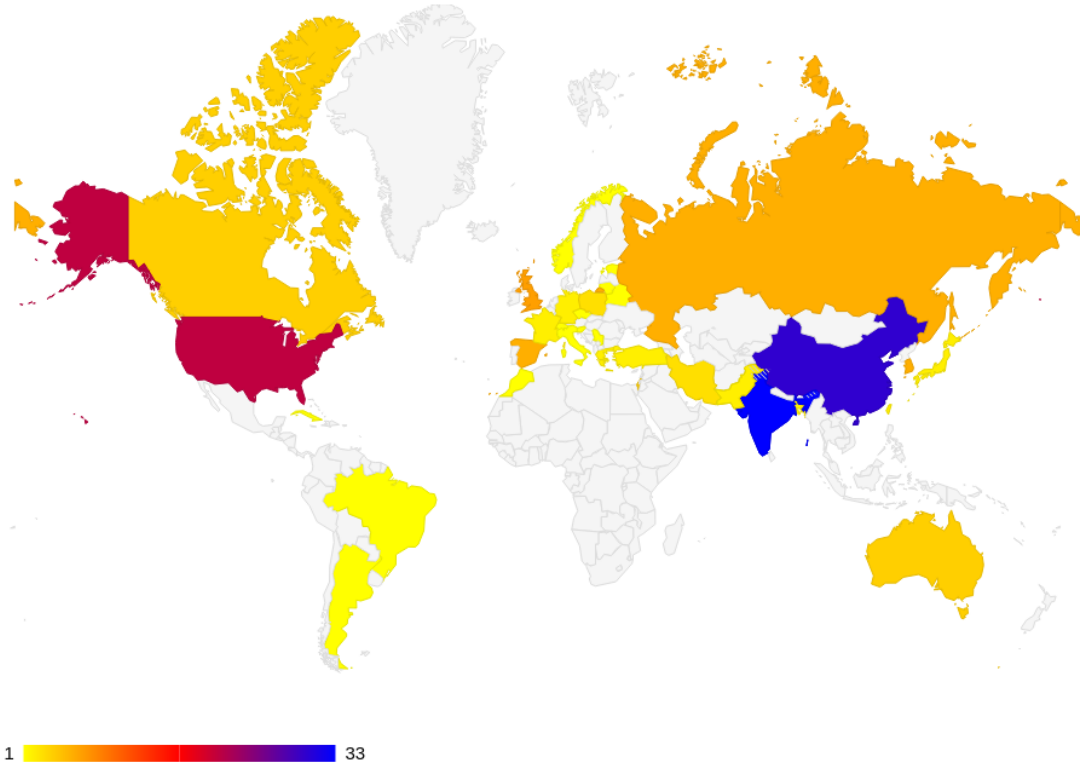
- **primary metric:** minimum normalized tandem Detection Cost Function (t-DCF) [1]
- **secondary metric:** equal error rate (EER) of *spoof* - *nonspoof* discrimination

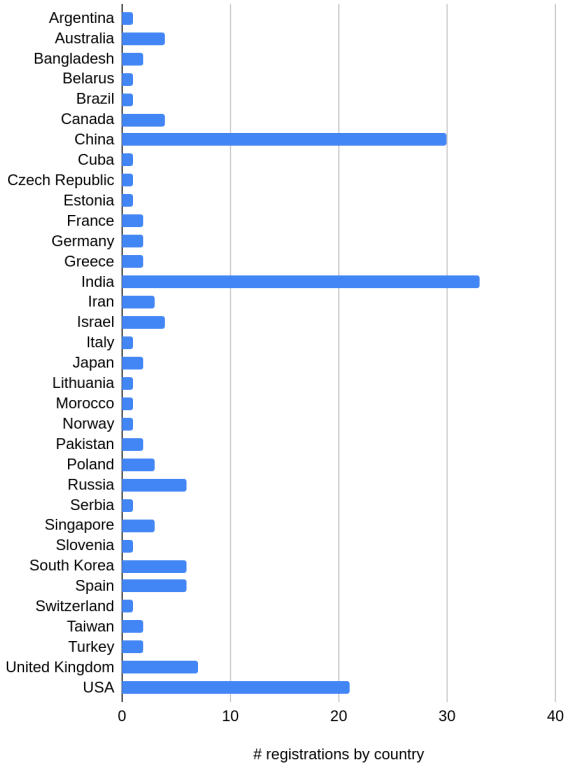
[1] T. Kinnunen, K. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in Proc. Odyssey, Les Sables d'Olonne, France, June 2018.

participant statistics

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







database & protocols

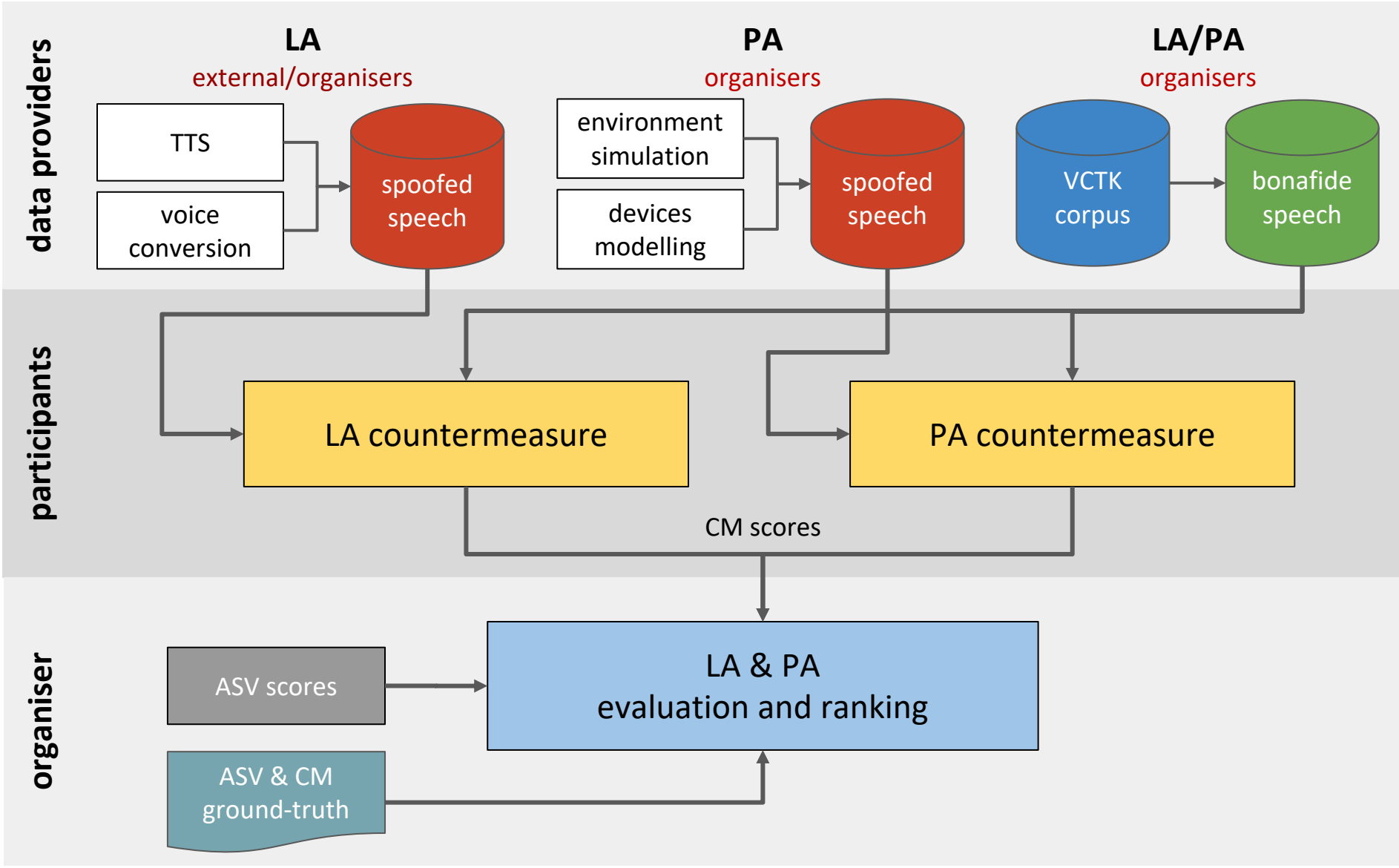
- based on **VCTK corpus** [1]
  - hemi-anechoic chamber of the University of Edinburgh
  - omni-directional head-mounted microphone (DPA 4035)
  - 96 kHz sampling frequency @ 24 bits
  - downsampled to 16 kHz @ 16 bits

- 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.

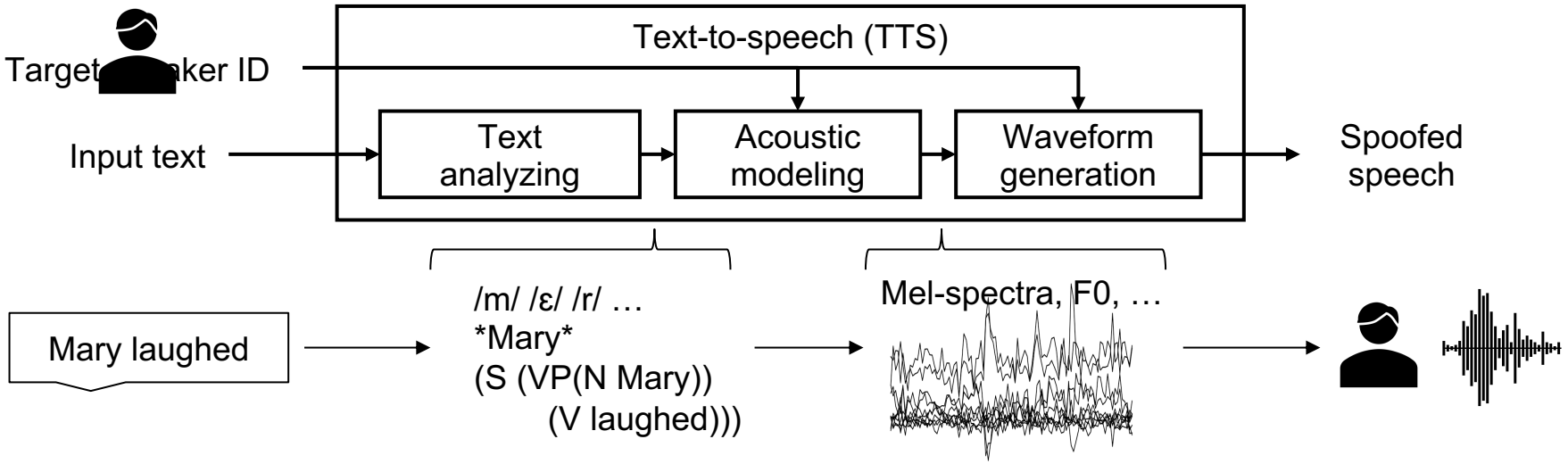
flow of the ASVspoof 2019 challenge

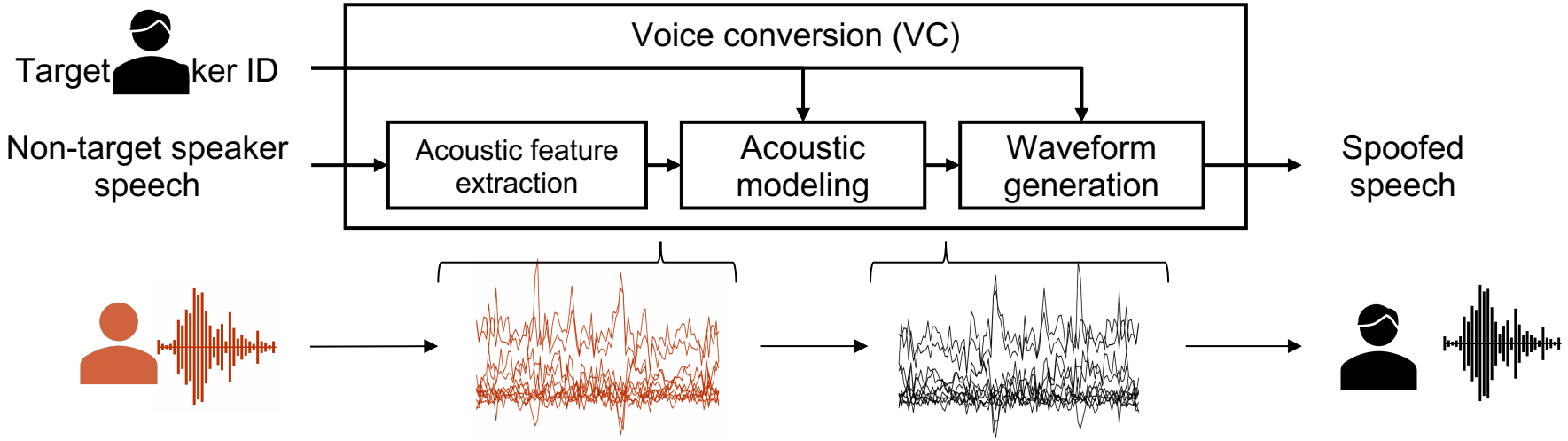


logical access

logical access

- Spoofing TTS/VC systems in general





logical access

- Spoofing TTS/VC systems in general

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
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

Varied or improved TTS/VC algorithms

logical access

- Spoofing TTS/VC systems

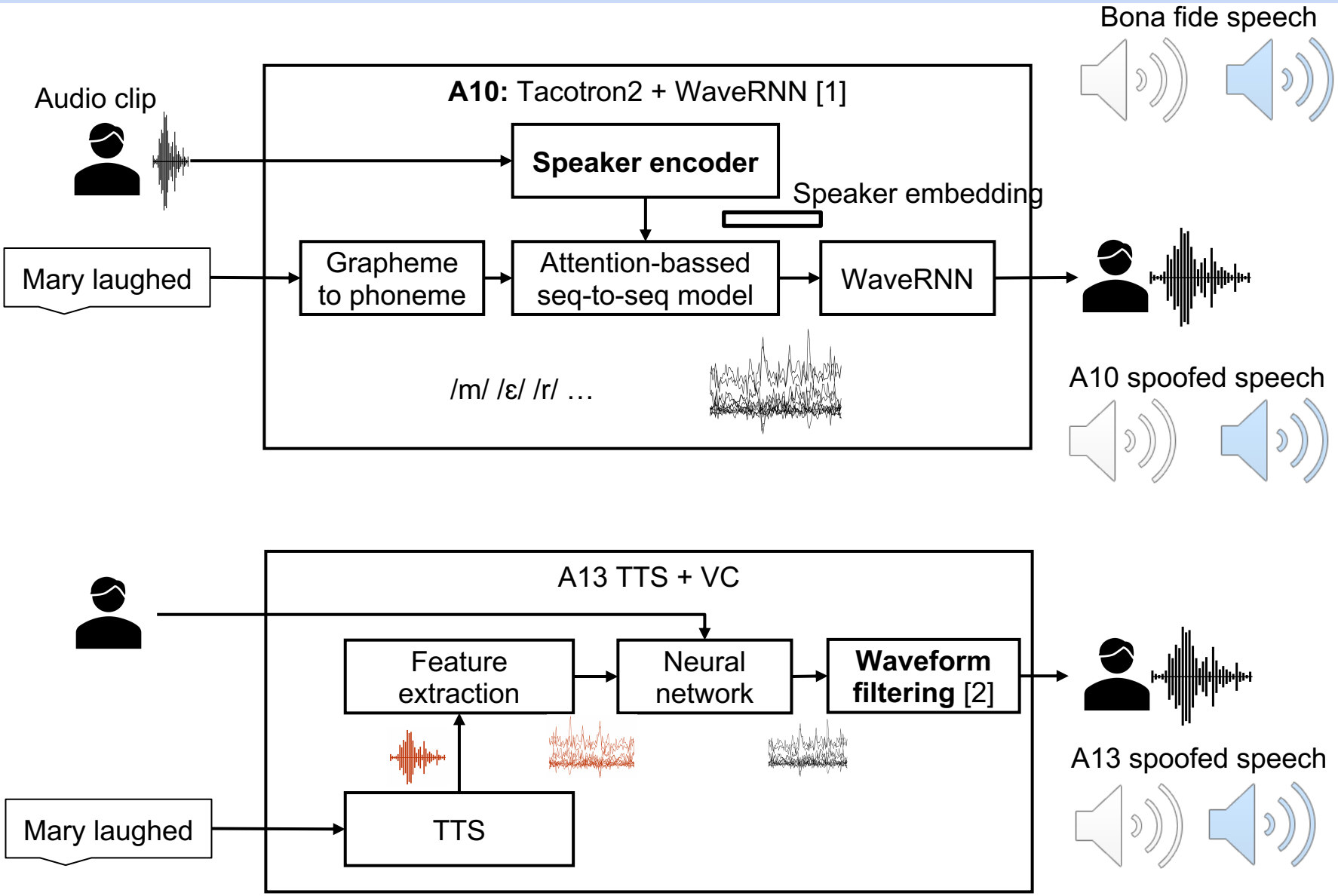
	Category	Acoustic model	Waveform generator	
A01	TTS	VAE + AR LSTM-RNN	WaveNet	Train & dev
A02	TTS	VAE + AR LSTM-RNN	WORLD	
A03	TTS	Feedforward NN	WORLD	
A04	TTS	-	Waveform concat.	
A05	VC	VAE	WORLD	
A06	VC	GMM-UBM	Spectral filtering	
A07	TTS	LSTM-RNN	WORLD + GAN	Evaluation
A08	TTS	AR LSTM-RNN	Neural source-filter model	
A09	TTS	LSTM-RNN	Vocaine	
A10	TTS	Attention seq2seq model	WaveRNN	Unknown TTS/VC algorithms
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	

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	0.0
	A02	TTS	VAE + AR LSTM-RNN	WORLD	15.04	0.0
	A03	TTS	Feedforward NN	WORLD	56.94	0.0
	A04	TTS	-	Waveform concat.	63.02	0.0
	A05	VC	VAE	WORLD	21.90	0.0
	A06	VC	GMM-UBM	Spectral filtering	10.11	0.0
Eva.	A07	TTS	LSTM-RNN	WORLD + GAN	59.68	0.02
	A08	TTS	AR LSTM-RNN	Neural source-filter model	40.39	0.09
	A09	TTS	LSTM-RNN	Vocaine	8.38	0.06
	A10	TTS	Attention seq2seq model	WaveRNN	57.73	12.21
	A11	TTS	Attention seq2seq model	Griffin-Lim	59.64	0.59
	A12	TTS	-	WaveNet	46.18	3.75
	A13	TTS-VC	Moment matching NN	Waveform filtering	46.78	12.42
	A14	TTS-VC	LSTM-RNN	STRAIGHT	64.01	2.88
	A15	TTS-VC	LSTM-RNN	WaveNet	58.85	3.22
	A16	TTS	-	Waveform concat.	64.52	0.02
	A17	VC	VAE	Waveform filtering	3.92	15.93
	A18	VC	i-vector/PLDA	MFCC-to-waveform	7.35	5.59
	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.

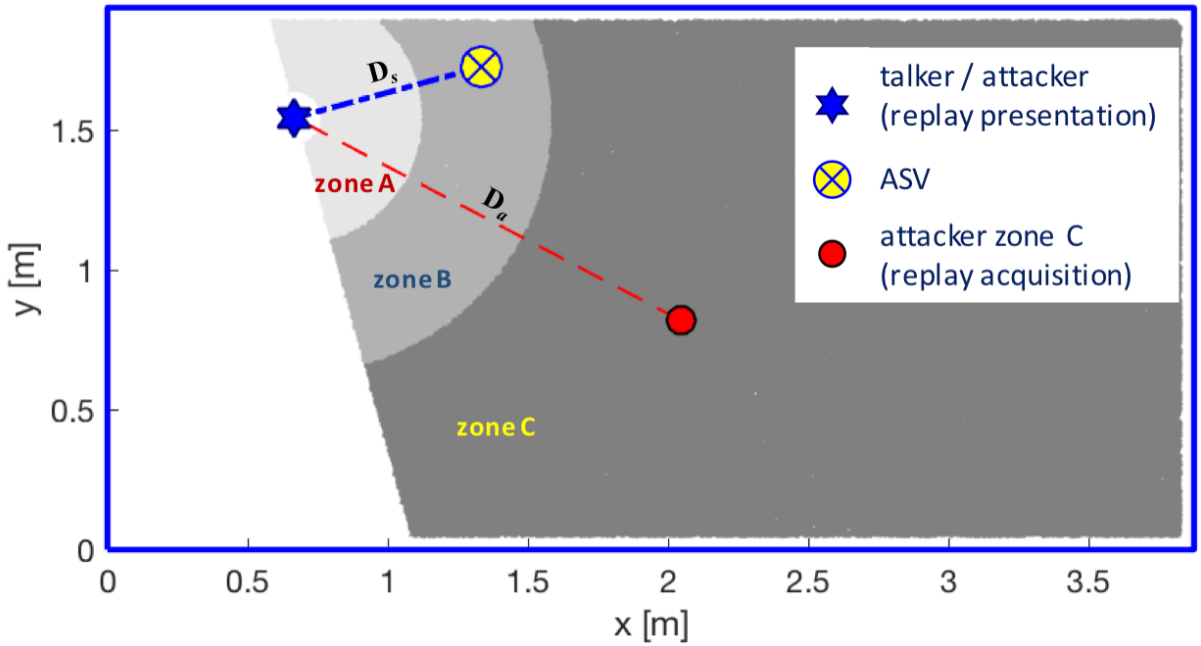
[2] E. Vincent. (2008) Roomsimove. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

[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) → categorical value
- **attack definition**
  - defined as duple (D_a, Q)
  - the set (A, B, C) → categorical value
- device quality (Q)
  - occupied bandwidth (OB) [kHz]
  - lower bound of OB (minF) [Hz]
  - linear/nonlinear OB power difference (linearity) [dB]

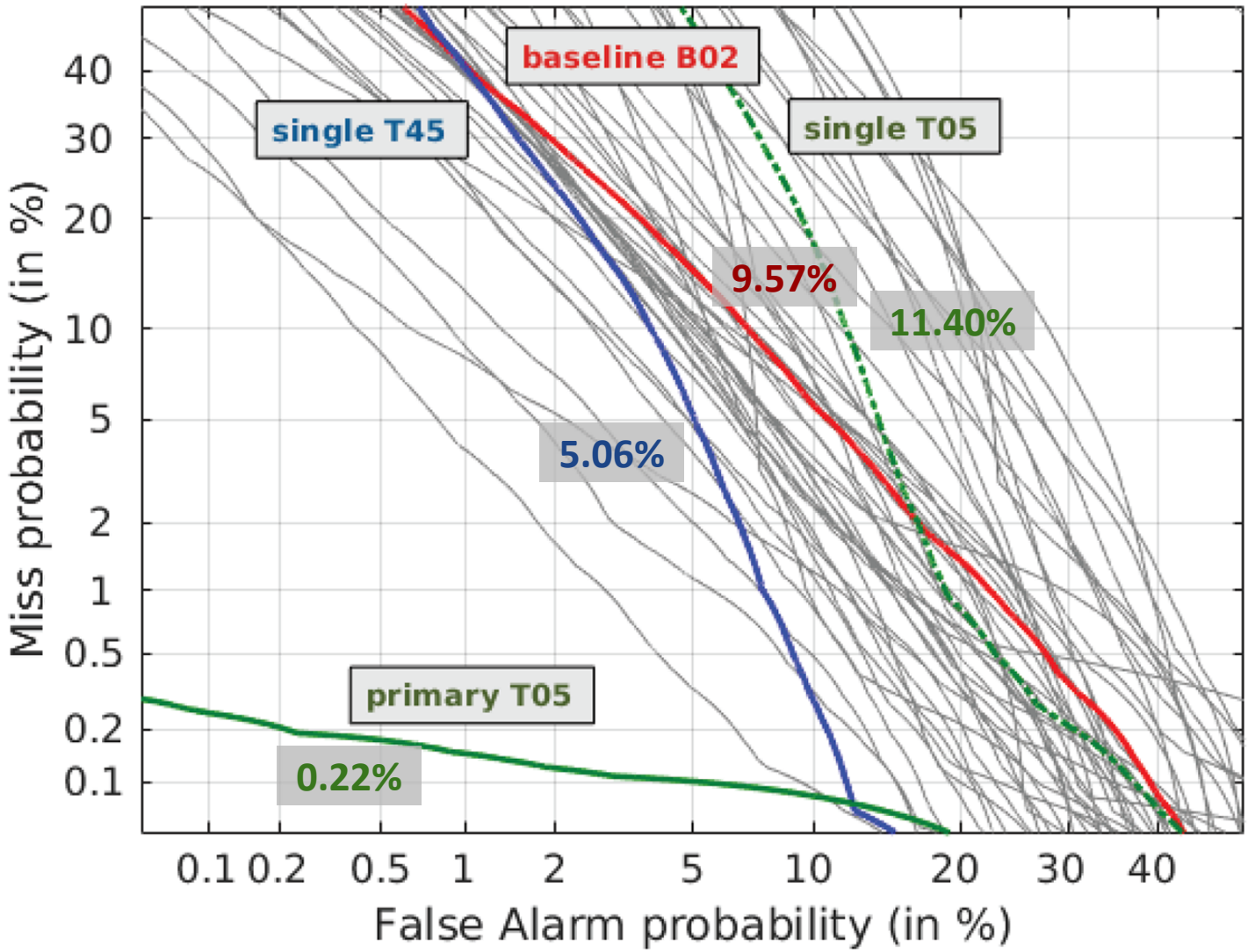
Environment definition	labels		
	a	b	c
S: Room size (square meters)	2-5	5-10	10-20
R: T60 (ms)	50-200	200-600	600-1000
D_s: Talker-to-ASV distance (cm)	10-50	50-100	100-150

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

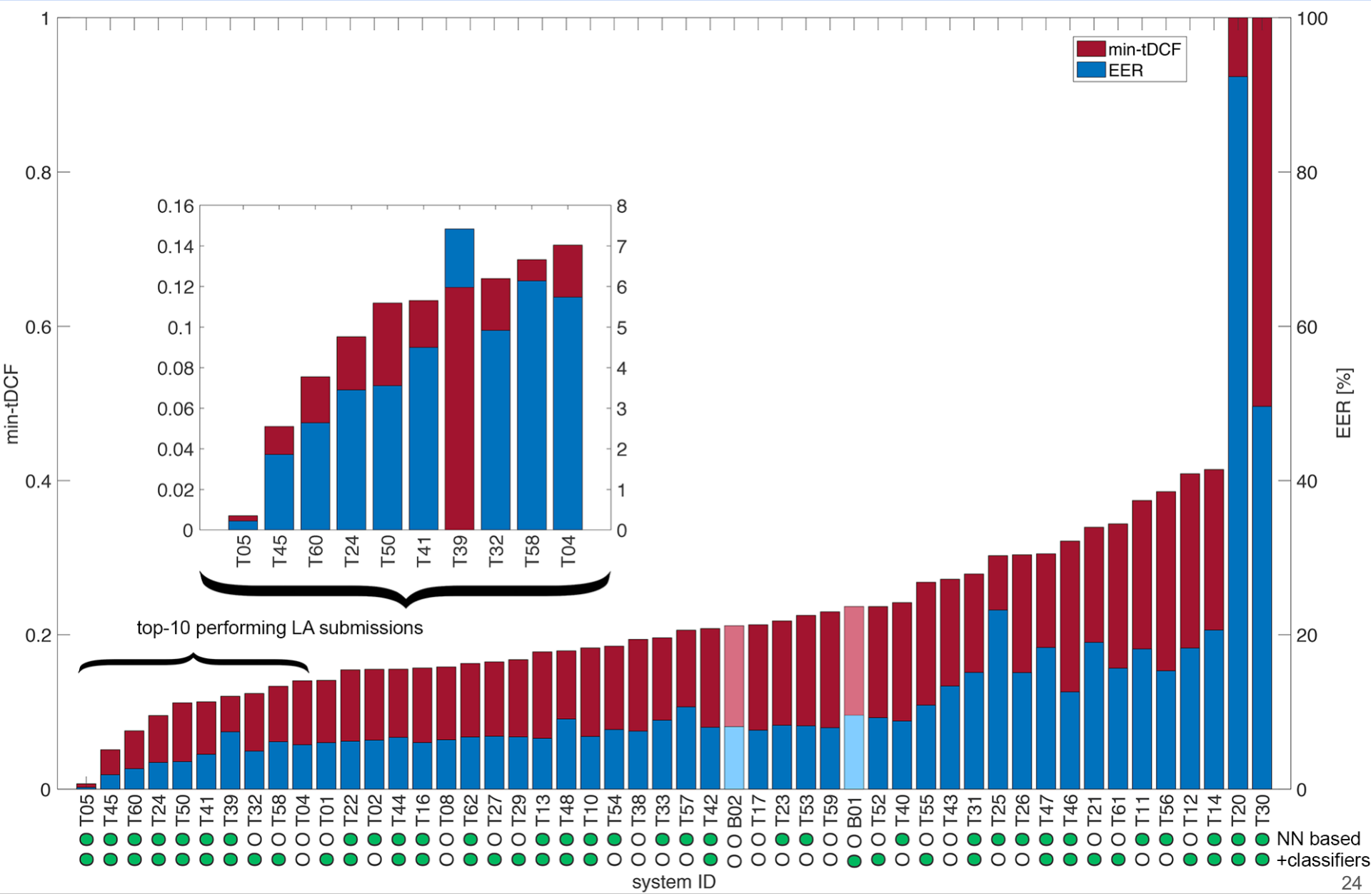
Replay device quality	OB (kHZ)	minF (Hz)	linearity (dB)
Perfect	inf	0	inf
High	> 10	< 600	> 100
Low	< 10	> 600	< 100

LA challenge results

LA - common primary submissions' results



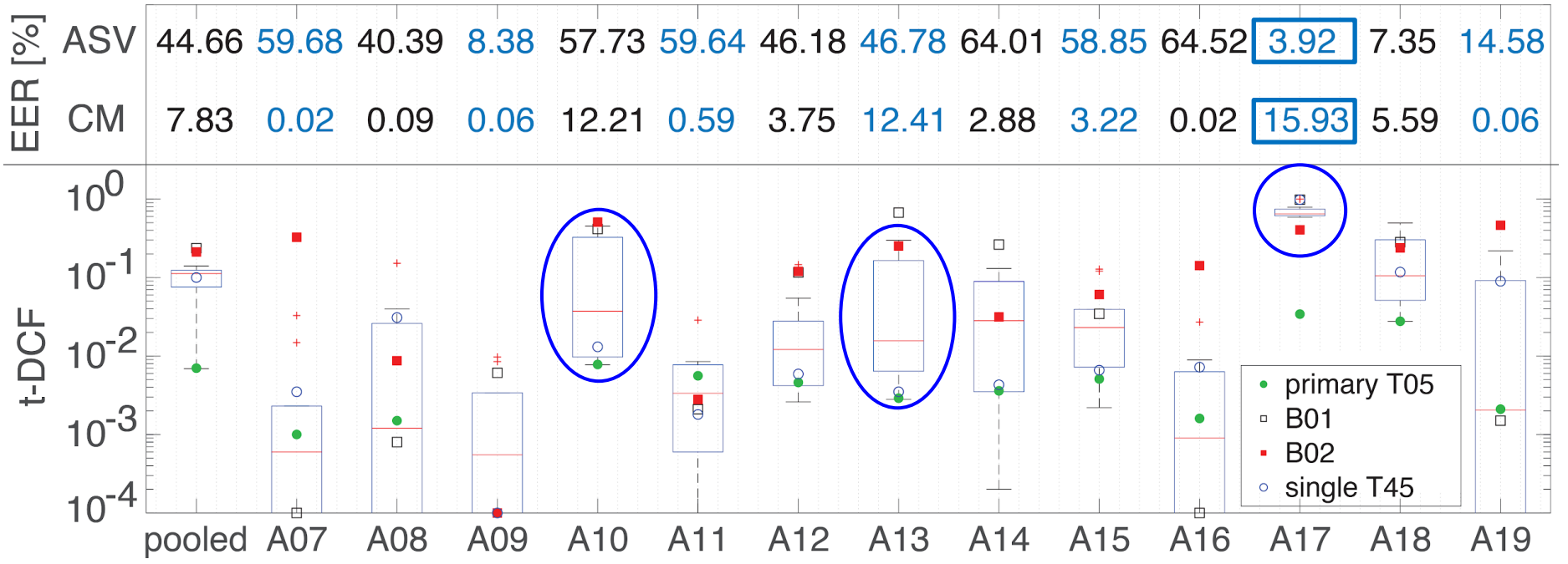
LA - common primary submissions' results

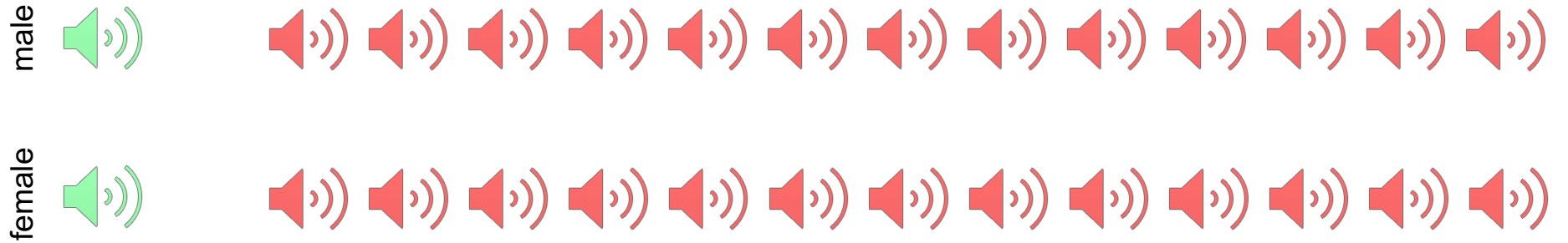


LA - 13 attacks breakdown

- ASV only zero-effort impostors → EER = 2.48%

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

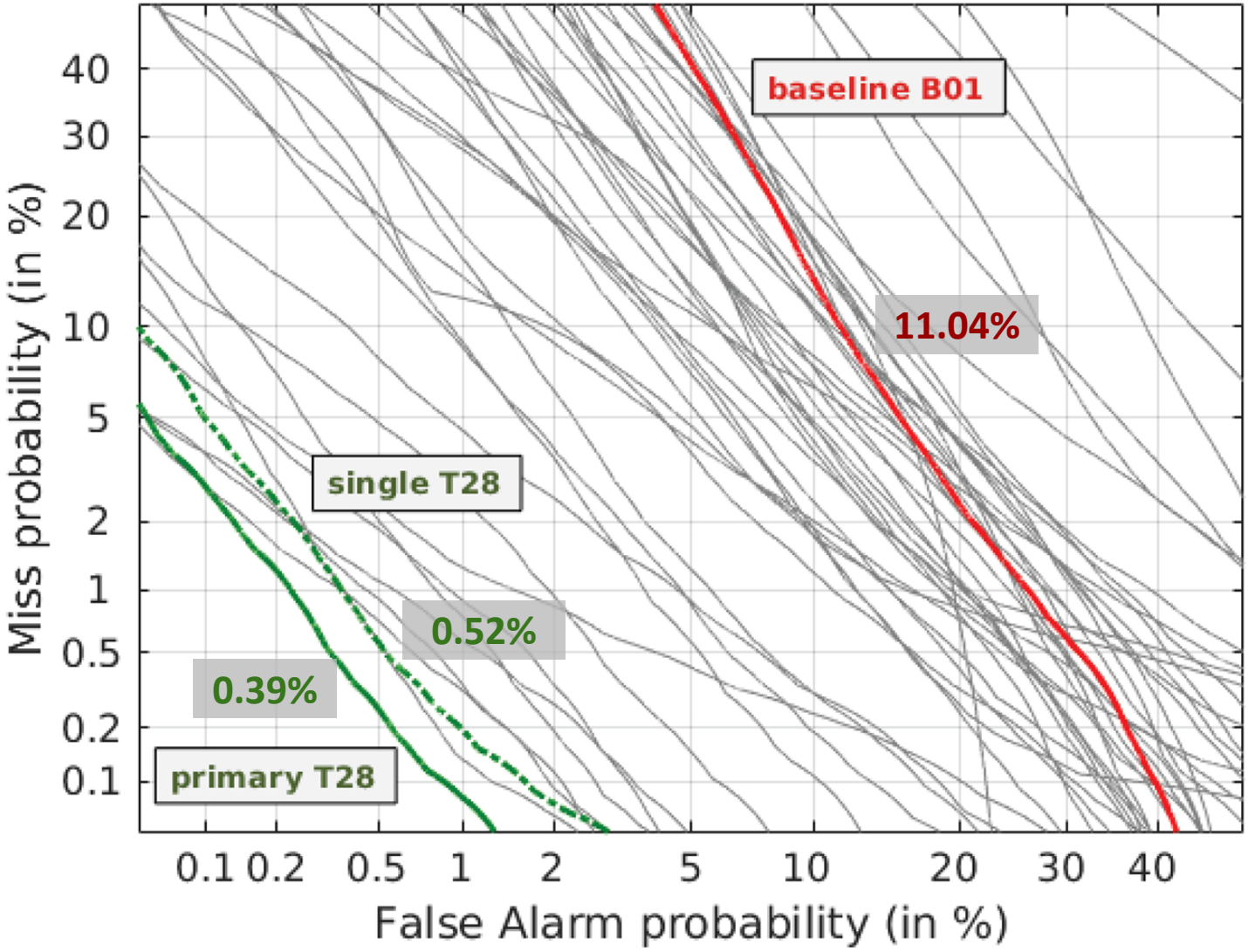




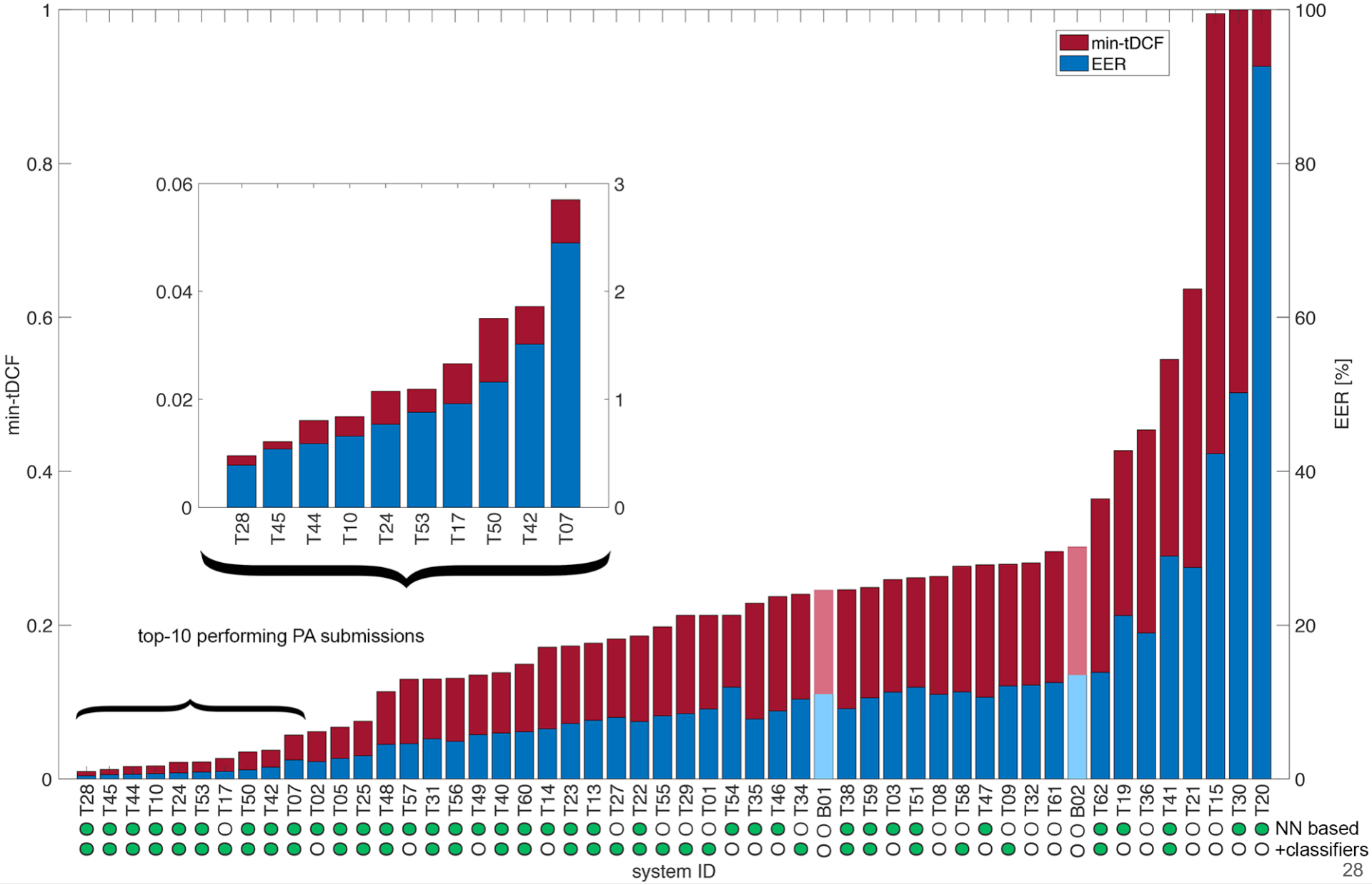
bonafide

PA challenge results

PA - common primary submissions' results



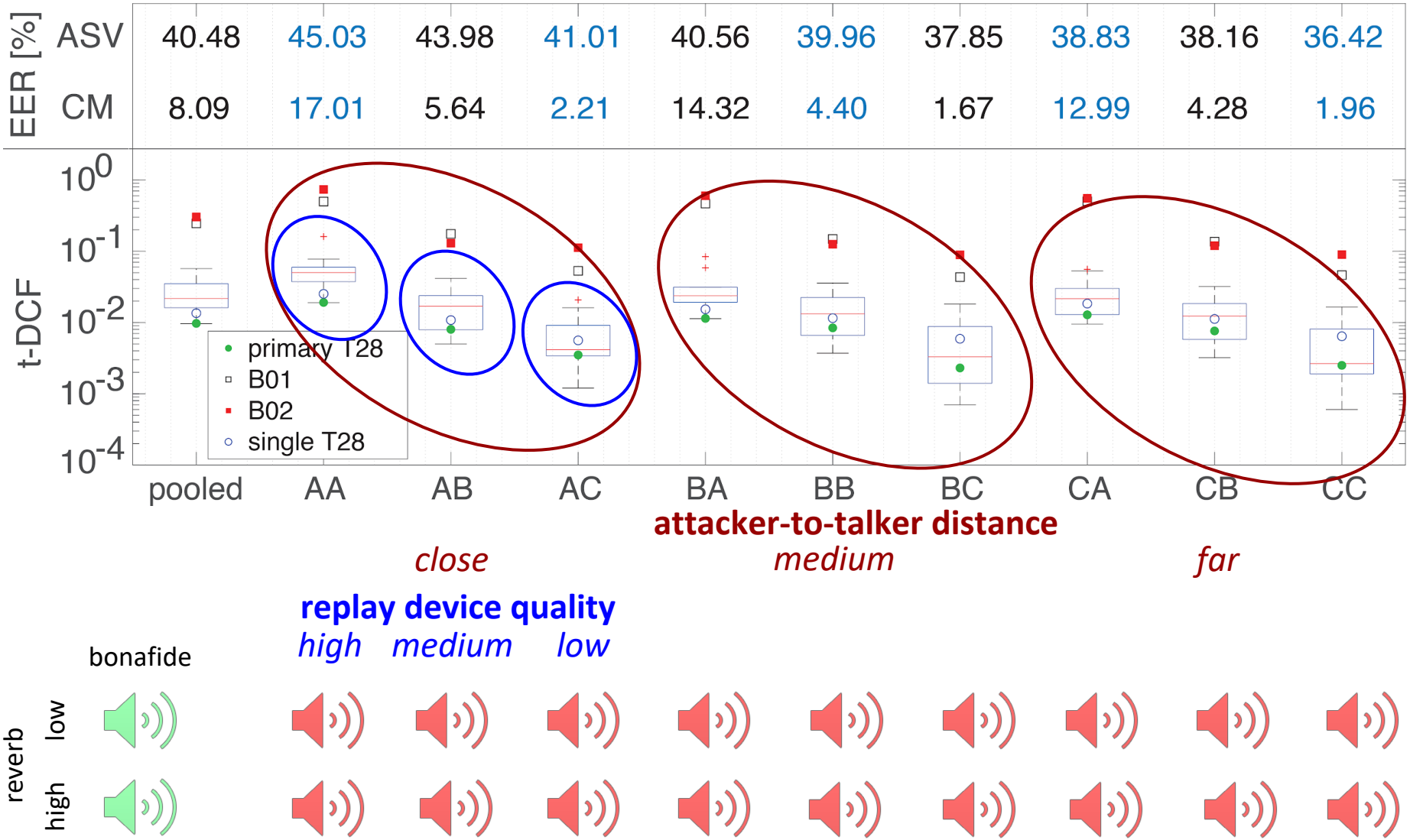
PA - common primary submissions' results



PA - 9 attacks breakdown

- ASV only zero-effort impostors → EER = 6.47%

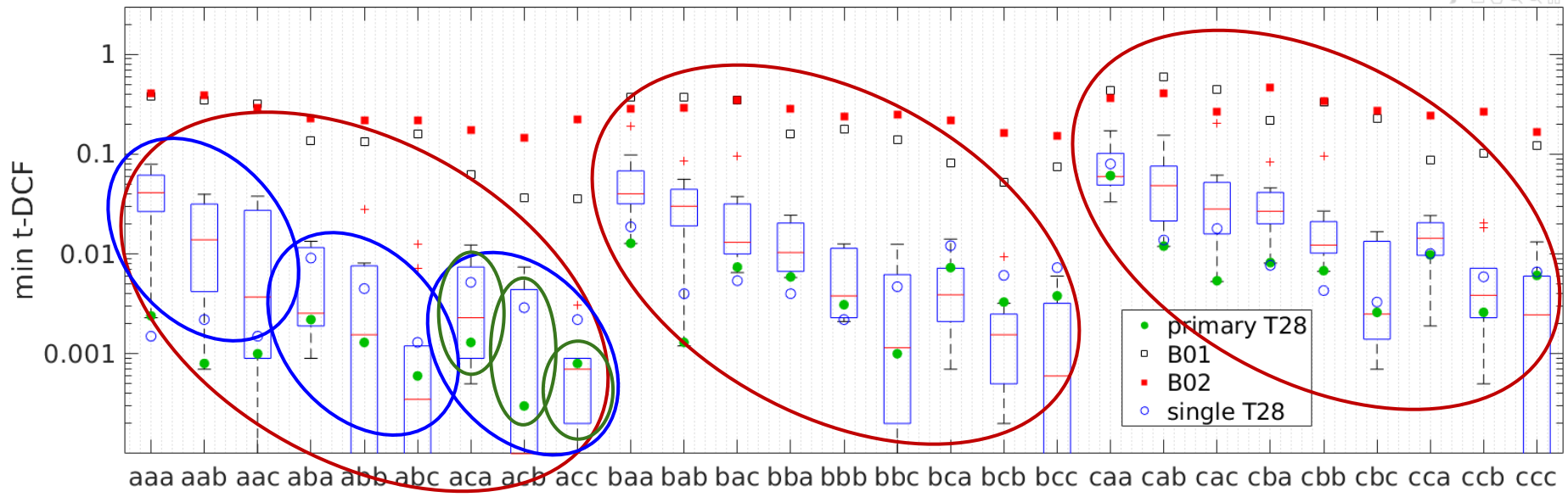
CM → median over all the primary submissions
t-DCF → top-10 primary submission



PA - 27 environments breakdown

EER [%]

	aaa	aab	aac	aba	abb	abc	aca	acb	acc	baa	bab	bac	bba	bbb	bbc	bca	bcb	bcc	caa	cab	cac	cba	cbb	cbc	cca	ccb	ccc
ASV	3.72	5.87	7.68	11.67	2.91	3.57	5.62	7.33	10.21	2.70	4.58	6.04	7.71	2.89	4.16	7.30	10.19	14.17	2.72	4.58	7.50	11.00	2.50	3.37	4.79	5.88	9.79
CM	9.11	4.03	3.45	1.34	10.98	9.69	4.02	3.08	2.01	14.24	6.78	4.63	3.06	12.11	7.91	3.38	1.92	1.56	11.33	4.53	3.66	2.00	13.28	12.74	6.36	3.22	2.81



room size

small *medium* *large*

reverberation noise

low *medium* *high*

talker-to-ASV distance

close med far

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

real PA

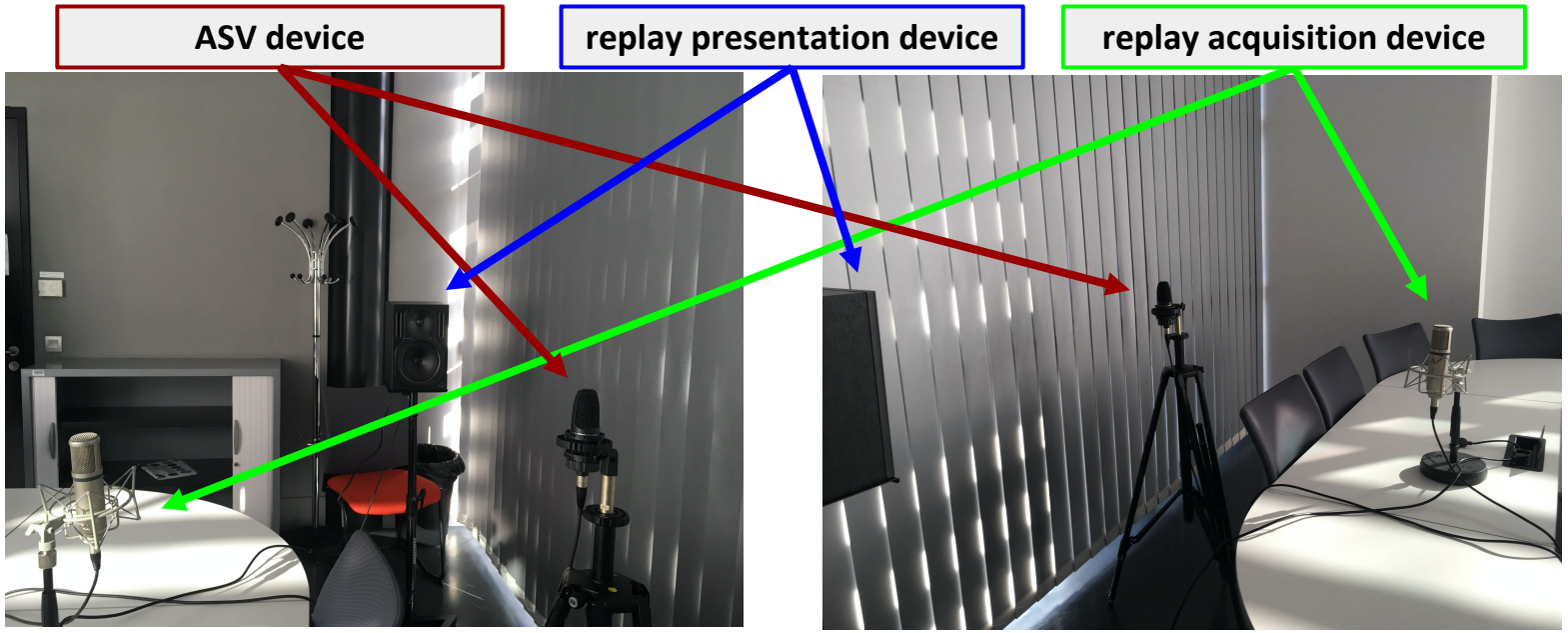
...hidden tracks in the album

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

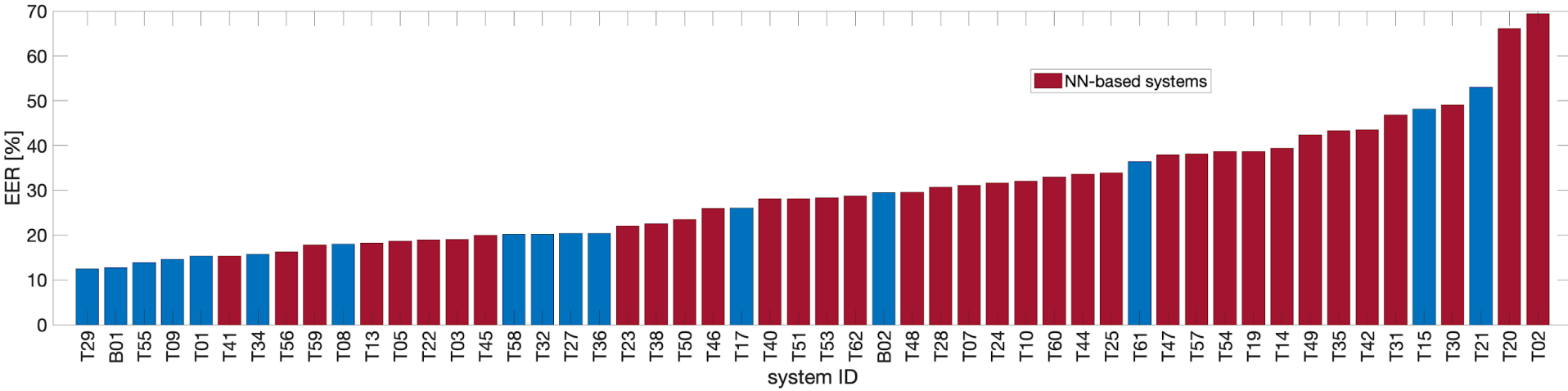
real PA

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

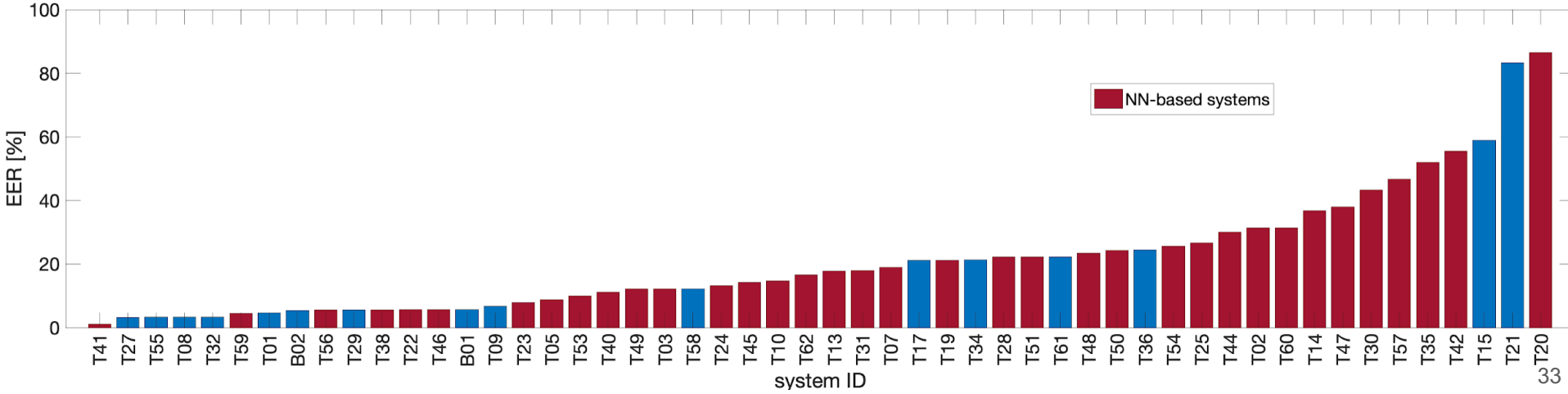


real PA - common primary submissions' results

- EER pooled over all attacks and labs



- single quietest lab (l1)
- 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