Workshop – 16th September 2021

ASVspoof

Automatic Speaker Verification and Spoofing Countermeasures Challenge

A satellite event of

30 August – 3 September 2021 INTERSPEECH 2021 BRNO | CZECHIA *Speech everywhere!*

ASVspoof 2021

Accelerating progress in spoofed and deepfake speech detection

Junichi Yamagishi, NII, Japan
Xin Wang, NII, Japan
Massimiliano Todisco, EURECOM, France
Md Sahidullah, Inria, France
Jose Patino, EURECOM, France
Andreas Nautsch, EURECOM, France
Xuechen Liu, University of Eastern Finland, Finland
Kong Aik Lee, Institute for Infocomm Research, Singapore
Tomi Kinnunen, University of Eastern Finland, Finland
Nicholas Evans, EURECOM, France
Héctor Delgado, Nuance Communications, Spain

EURECOM Sophia Antipolis · NII National Institute of Informatics · UNIVERSITY OF EASTERN FINLAND · Inria · Institute for Infocomm Research I²R

Outline

Introduction

Logical access

Physical access

Speech deepfake

Database

Unseen conditions

Results

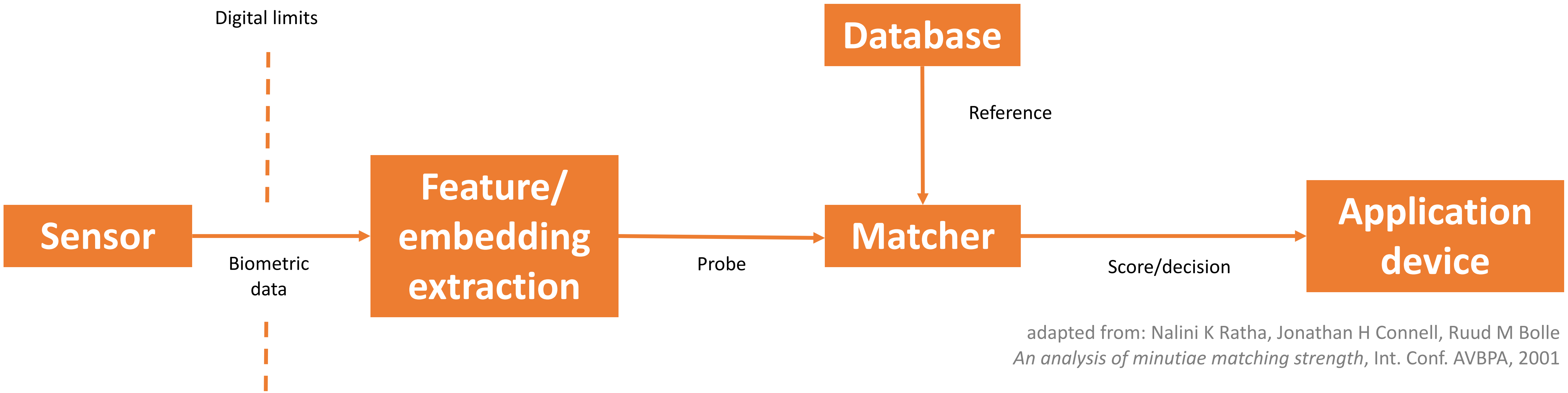
Hidden tracks

Conclusions

Security / spoofing in voice biometrics

- persons masquerading as others in order to gain illegitimate access to sensitive or protected resources
- a.k.a. presentation attacks (ISO / IEC)

ISO/IEC 30107-1
http://standards.iso.org/ittf/PubliclyAvailableStandards/c053227_ISO_IEC_30107-1_2016.zip



adapted from: Nalini K Ratha, Jonathan H Connell, Ruud M Bolle
An analysis of minutiae matching strength, Int. Conf. AVBPA, 2001

- attacks either pre-sensor (PA) or post-sensor (LA)

In the media


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


BBC | News | Sport | Weather | Shop | Earth | Travel | Capital | More | Search

NEWS

Home | Video | World | UK | Business | Tech | Science | Magazine | Entertainment & Arts | Health | More

Technology

Adobe Voco 'Photoshop-for-voice' causes concern

7 November 2016 | Technology

Share

ADOBE

Adobe has yet to say when it might release Voco to the public

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


UAB News

Knowledge that will change your world

search UAB News | go

Public Relations: (205) 934-3884 | UAB Quicklinks

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 Shonesy

September 25, 2015 | Print | Email

University of Alabama at Birmingham researchers have found that automated and human verification for voice-based user authentication systems are vulnerable to voice impersonation attacks. 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-morphing tool, the researchers developed a voice impersonation attack to attempt to penetrate automated and human verification systems.

Click to enlarge


The Telegraph | ALL SECTIONS

Technology | More

Technology

This robot speech simulator can imitate anyone's voice

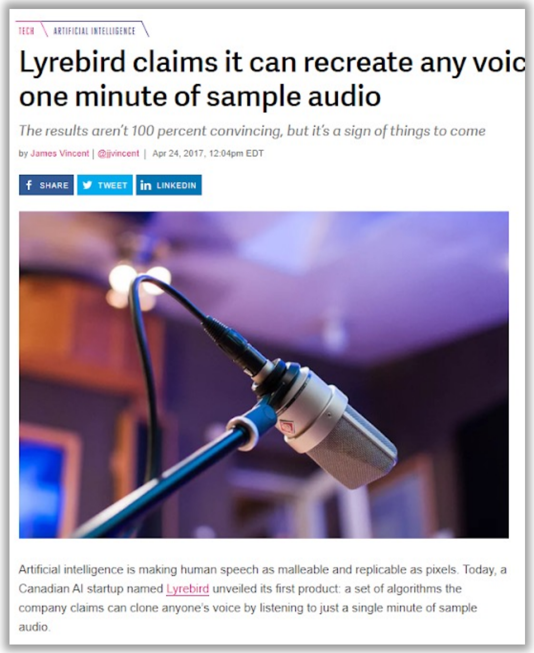
share | 0 Comments

The machine has mimicked Barack Obama CREDIT: REX


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

HSBC said it is to review security on its voice-access systems following the breach. Photograph: Stefan Wermuth/Reuters


TECH \ ARTIFICIAL INTELLIGENCE

Lyrebird claims it can recreate any voic one minute of sample audio

The results aren't 100 percent convincing, but it's a sign of things to come

by James Vincent | @jjvincent | Apr 24, 2017, 12:04pm EDT

SHARE | TWEET | LINKEDIN

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.


Analyzing The Rise Of Deepfake Voice Technology

Dominic David Forbes Councils Member

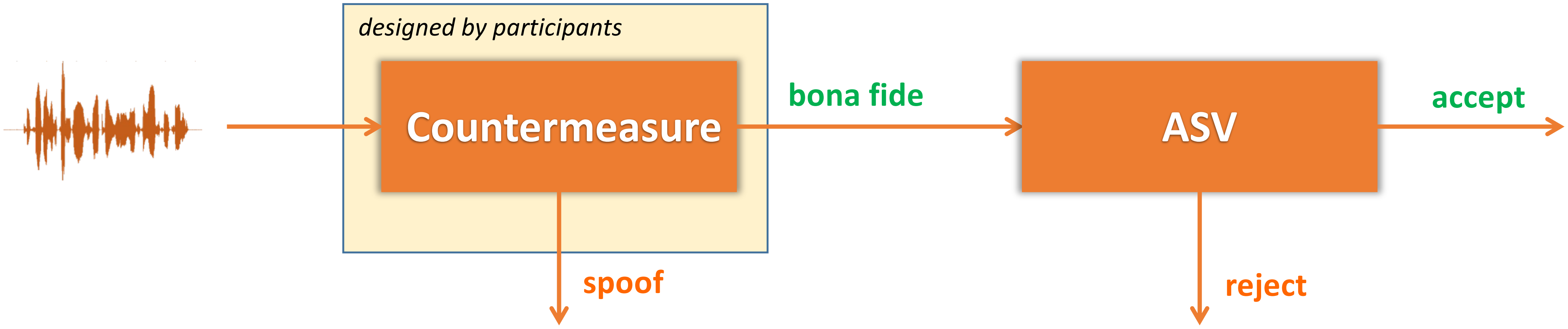
Forbes Technology Council COUNCIL POST | Membership (fee-based)

Innovation

Chief Executive of ACTIONX — leaders in data science and artificial intelligence.

ASVspoof

- spoofing detection task with an optional focus upon ASV



- common databases, protocols and metrics
  - tandem detection cost function (t-DCF) [1] – LA, PA
  - equal error rate (EER) – DF
- ASV system based on ASVTorch [2] + Kaldi [3]
- 4 (unoptimised) countermeasure baselines:
  - CQCC-GMM, LFCC-GMM, LFCC-LCNN, RawNet2

[1] T. Kinnunen et al. "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification". In Proc. *Speaker Odyssey* , 2018
[2] K. A. Lee et al. "ASVtorch toolkit: Speaker verification with deep neural networks". *SoftwareX* 14, 2021
[3] D. Povey et al. "The Kaldi speech recognition toolkit". In Proc. *IEEE ASRU*, 2011

Progress

First common datasets, metrics and protocols; modest attack variability

ASVspoof initiative launched

ASVspoof 2017 PA

Small, purpose collected datasets; low attack variability

1999 | 2013 | 2014 | 2015 | 2017 | 2019 | 2021

First special session @ INTERSPEECH 2013

ASVspoof 2015 LA

Tandem assessment

ASVspoof 2019 LA + PA

Large, standard datasets adapted to study spoofing; low attack variability

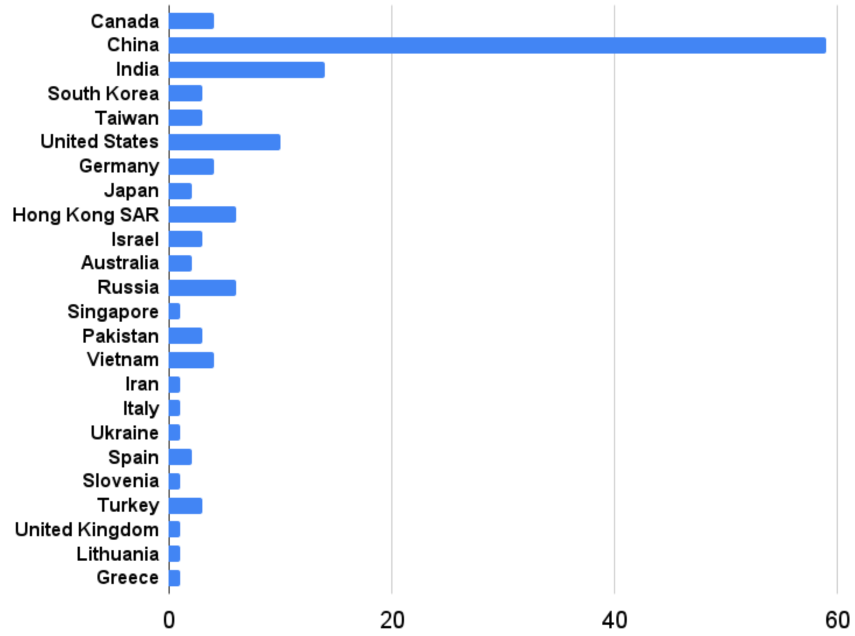
Improved methodology; broader attack variability

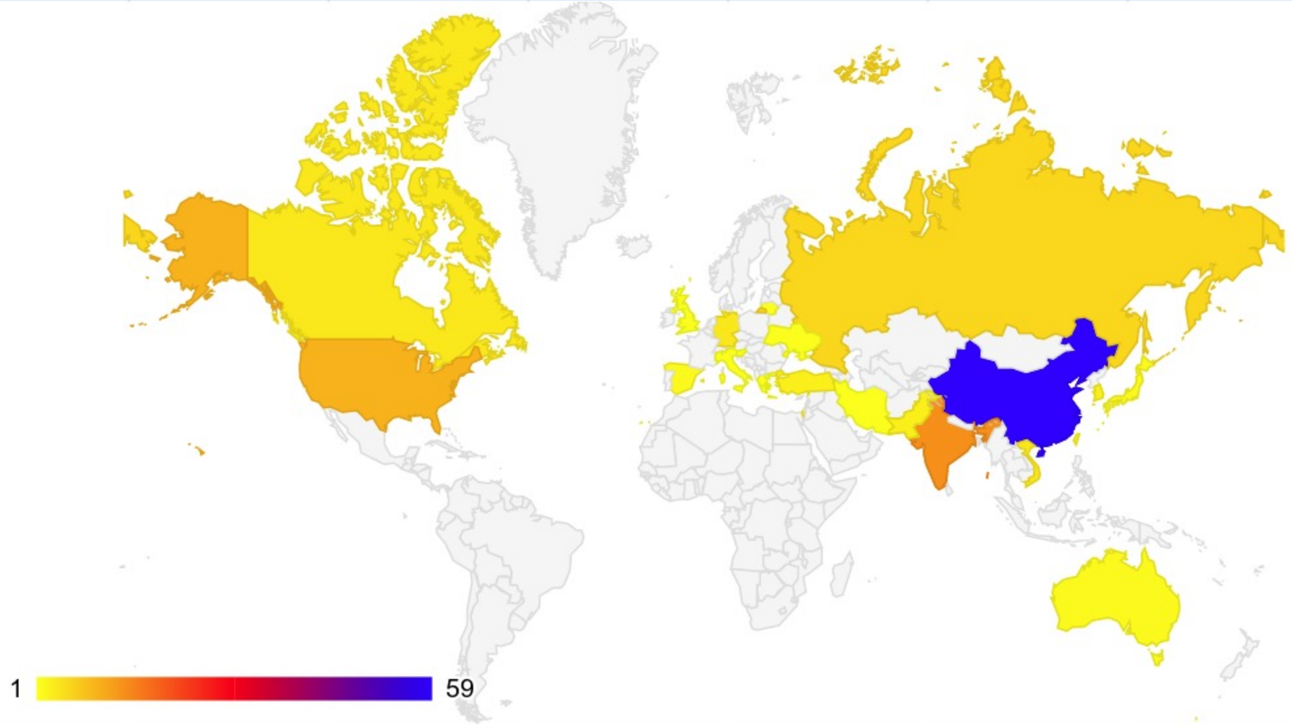
Neural & acoustic waveform models; controlled replay scenario

- LA: telephony encoding and transmission
- PA: real physical environments, variety of recording and replay devices
- DF: new speech deepfake detection task
- new evaluation platform, schedule & format

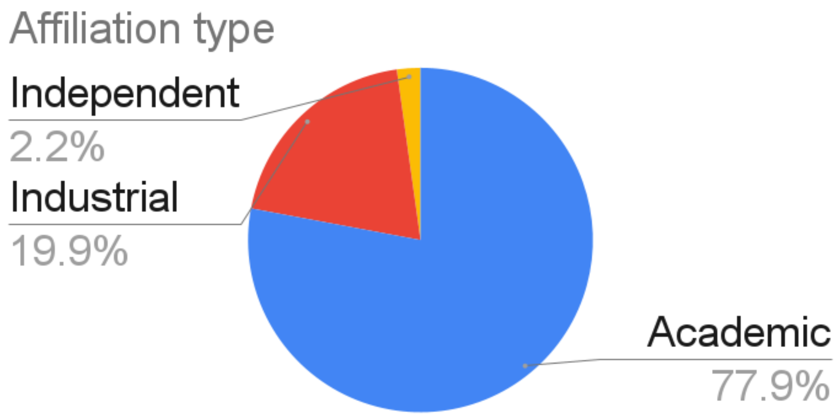
ASVspoof
Automatic Speaker Verification and Spoofing Countermeasures Challenge

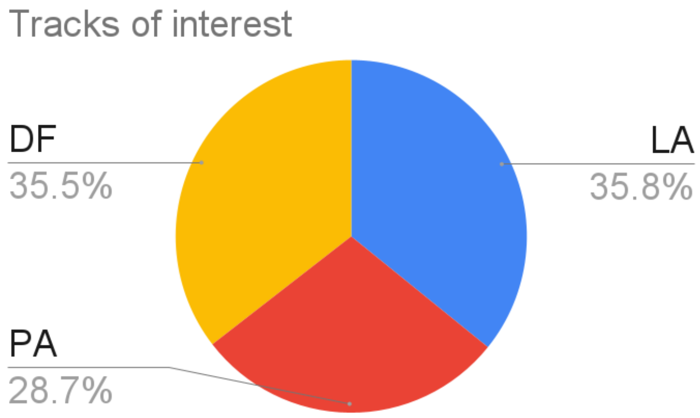
Challenge participant statistics

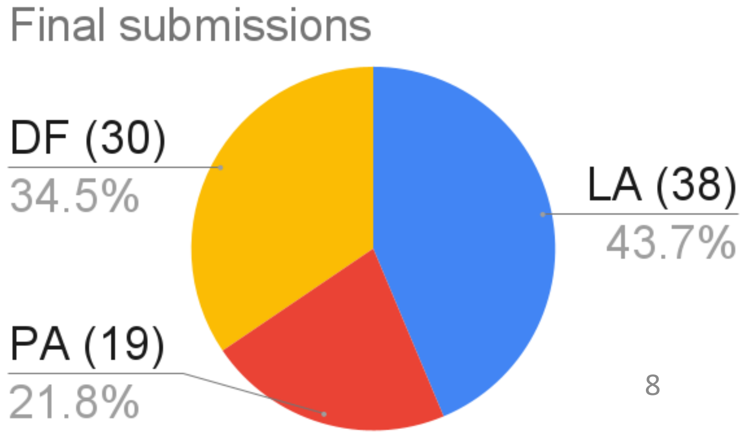




Registrations by country (137 total)

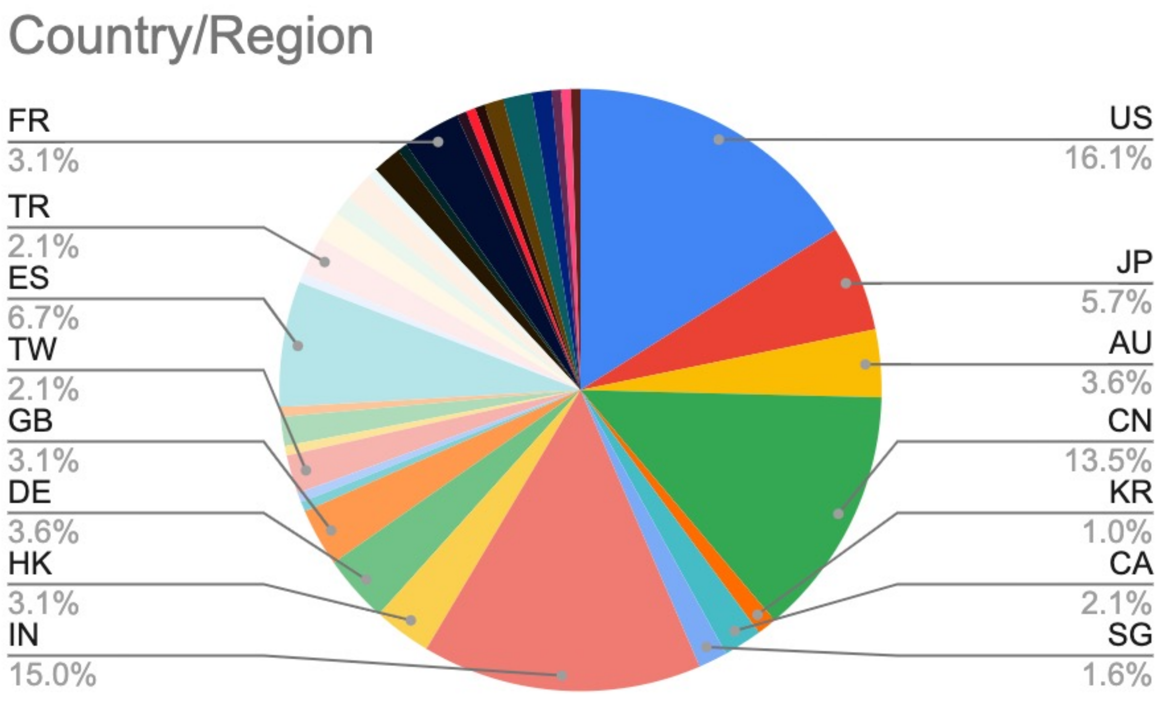


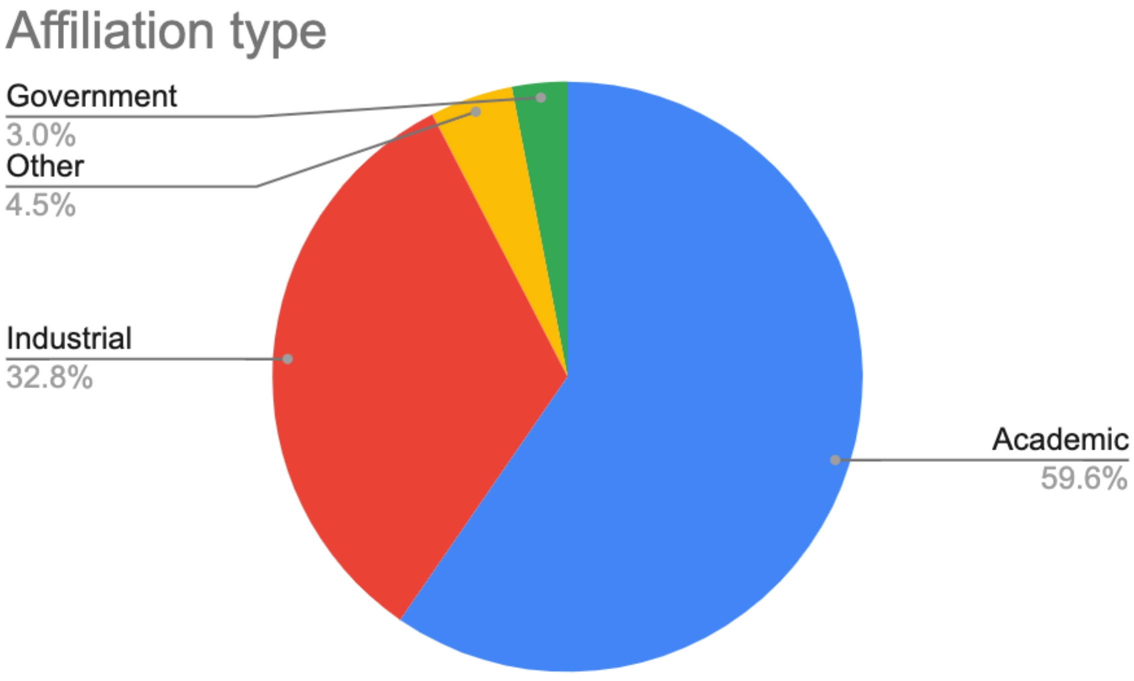




Workshop

- Registered attendees: **198** (by 2021-09-15)

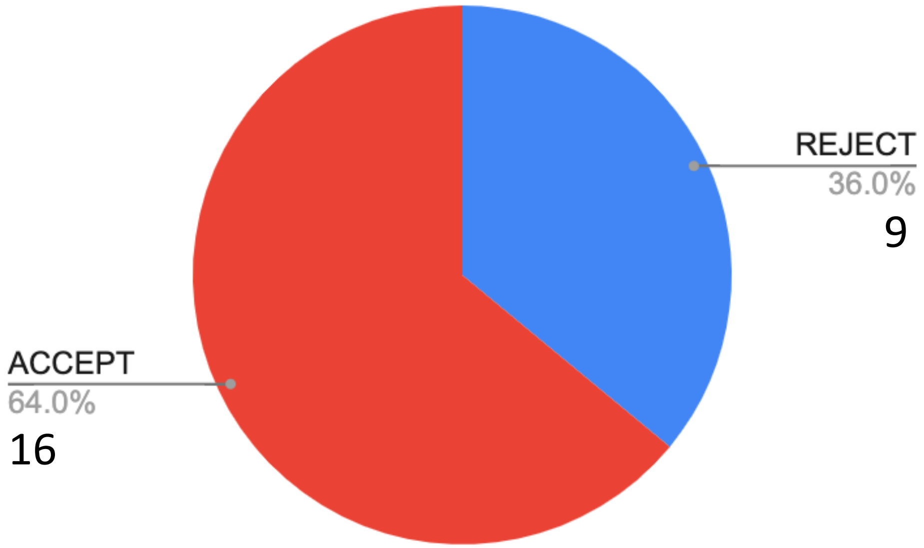




Workshop

- Papers
  - Submitted papers: 25
  - Acceptance rate: 64%

Challenge summary: 1
ASVspoof 2021 system descriptions: 10
Research papers on anti-spoofing: 5



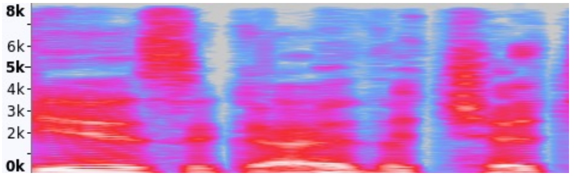
- Program: https://www.asvspoof.org/workshop
  - Time-zone of the presenters are considered
  - Each session mixes ASVspoof 2021 system descriptions and other research papers

Logical access (LA)

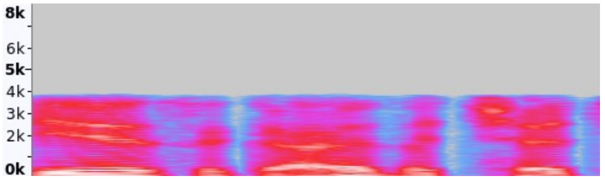
Logical access

- **New to 2021:** telephony encoding and transmission
- Real transmitted synthetic / converted speech
  - VoIP and PSTN setups
  - Sourced from ASVspoof 2019
    - Seen and unseen eval. partitions
- Codec effects: bandwidth, bitrate, ...
- Multi-site transmission setup

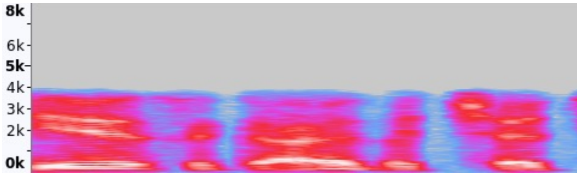
	Bona fide	Spoof
Progress phase	1,676	14,788
Evaluation phase	14,816	133,360
Sum	16,492	148,148
Hidden track	1,960	14,966



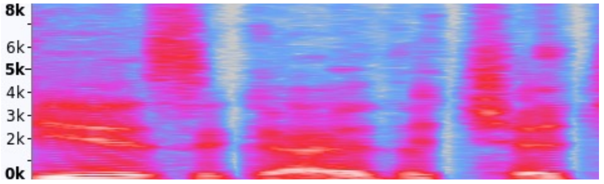
LA-C1



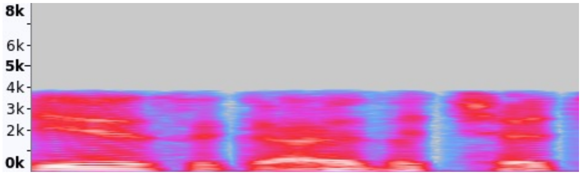
LA-C2



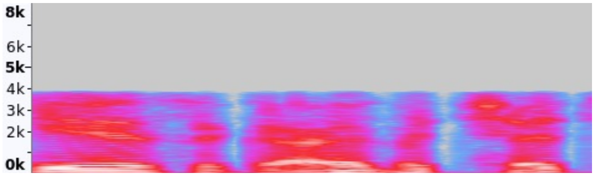
LA-C3



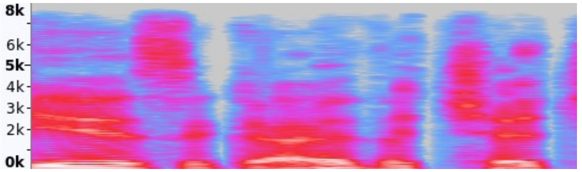
LA-C4



LA-C5



LA-C6



LA-C7

LA – Data collection

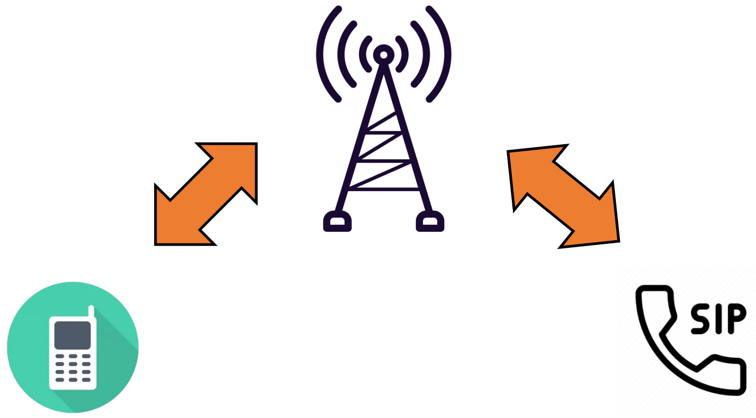
- VoIP setup:
  - Move towards a more realistic telephony scenario
  - Data collection using Asterisk PBX
  - Encoding and transmission

Cond.	Codec	Audio bandwitdh	Transmission	Bitrate
LA-C1	-	16 kHz	-	250 kbps
LA-C2	a-law	8 kHz	VoIP	64 kbps
LA-C3	unk. + μ-law	8 kHz	PSTN + VoIP	- / 64 kbps
LA-C4	G.722	16 kHz	VoIP	64 kbps
LA-C5	μ-law	8 kHz	VoIP	64 kbps
LA-C6	GSM	8 kHz	VoIP	13 kbps
LA-C7	OPUS	16 kHz	VoIP	VBR $\sim$16 kbps

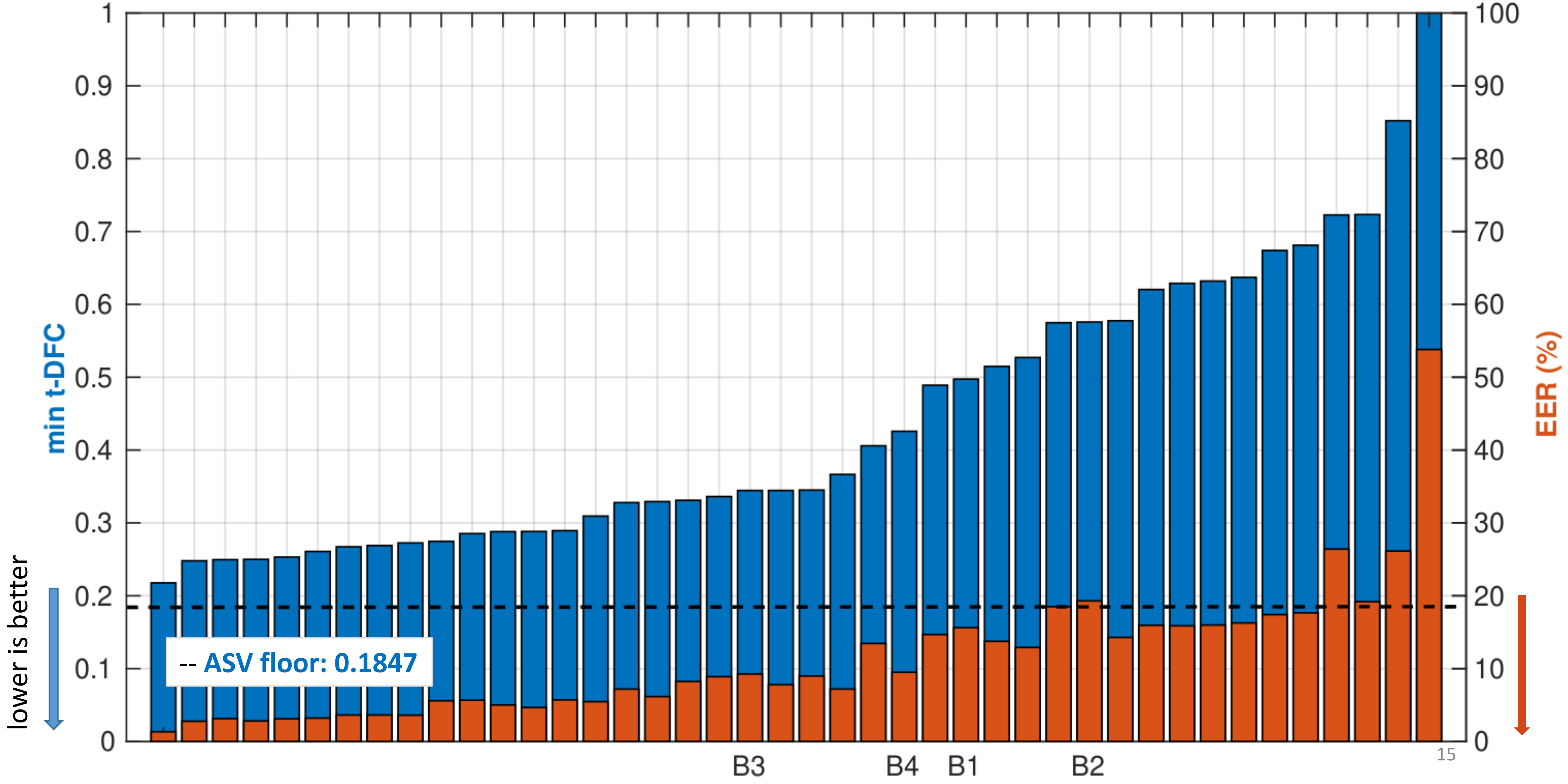
LA – Data collection

- PSTN + VoIP setup:
  - Real cellphone to SIP phone transmission
  - Certain transmission factors out of our control

Cond.	Codec	Audio bandwitdh	Transmission	Bitrate
LA-C1	-	16 kHz	-	250 kbps
LA-C2	a-law	8 kHz	VoIP	64 kbps
LA-C3	unk. + μ-law	8 kHz	PSTN + VoIP	- / 64 kbps
LA-C4	G.722	16 kHz	VoIP	64 kbps
LA-C5	μ-law	8 kHz	VoIP	64 kbps
LA-C6	GSM	8 kHz	VoIP	13 kbps
LA-C7	OPUS	16 kHz	VoIP	VBR $\sim$16 kbps

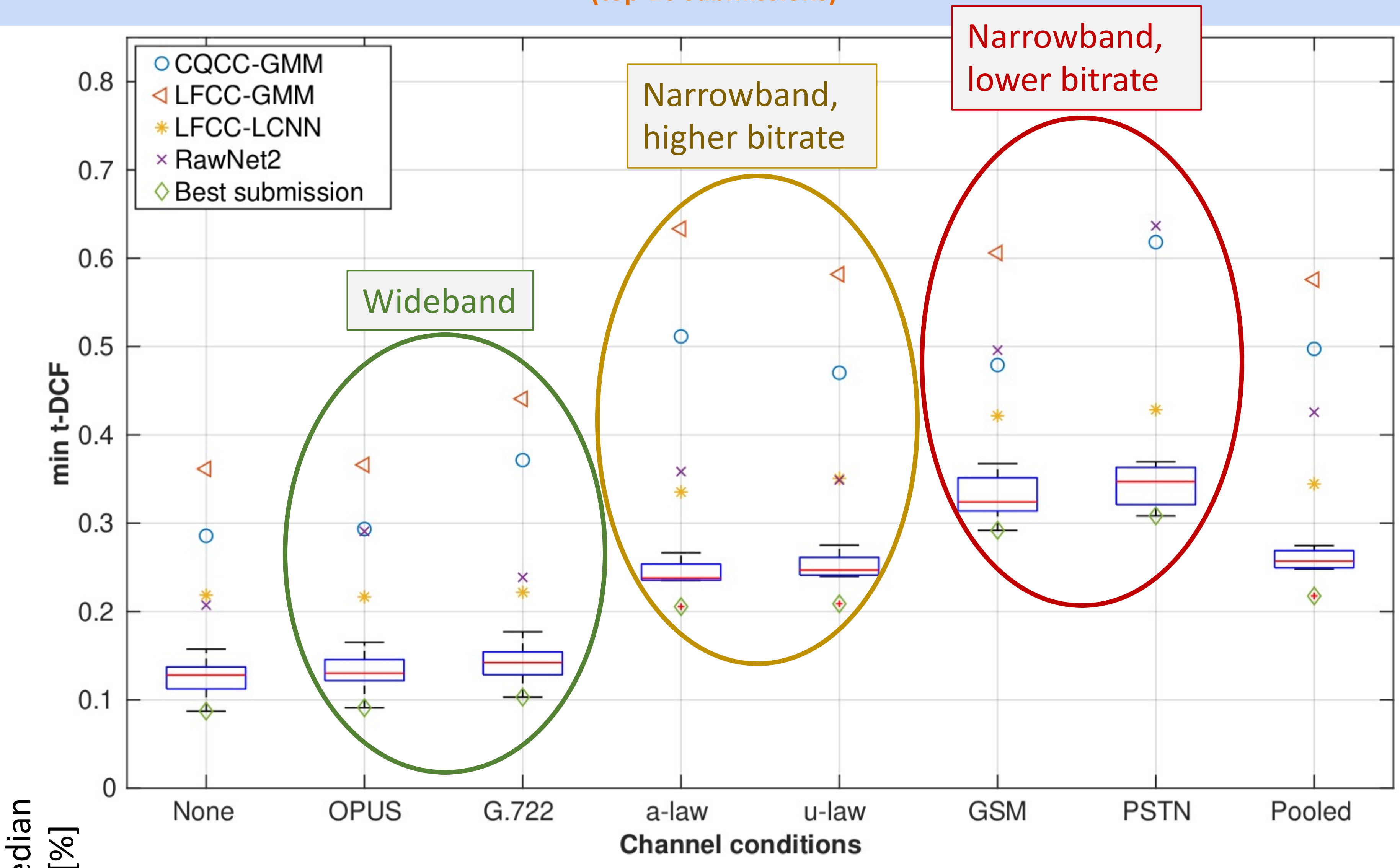


LA results: ranking



LA results: channel conditions

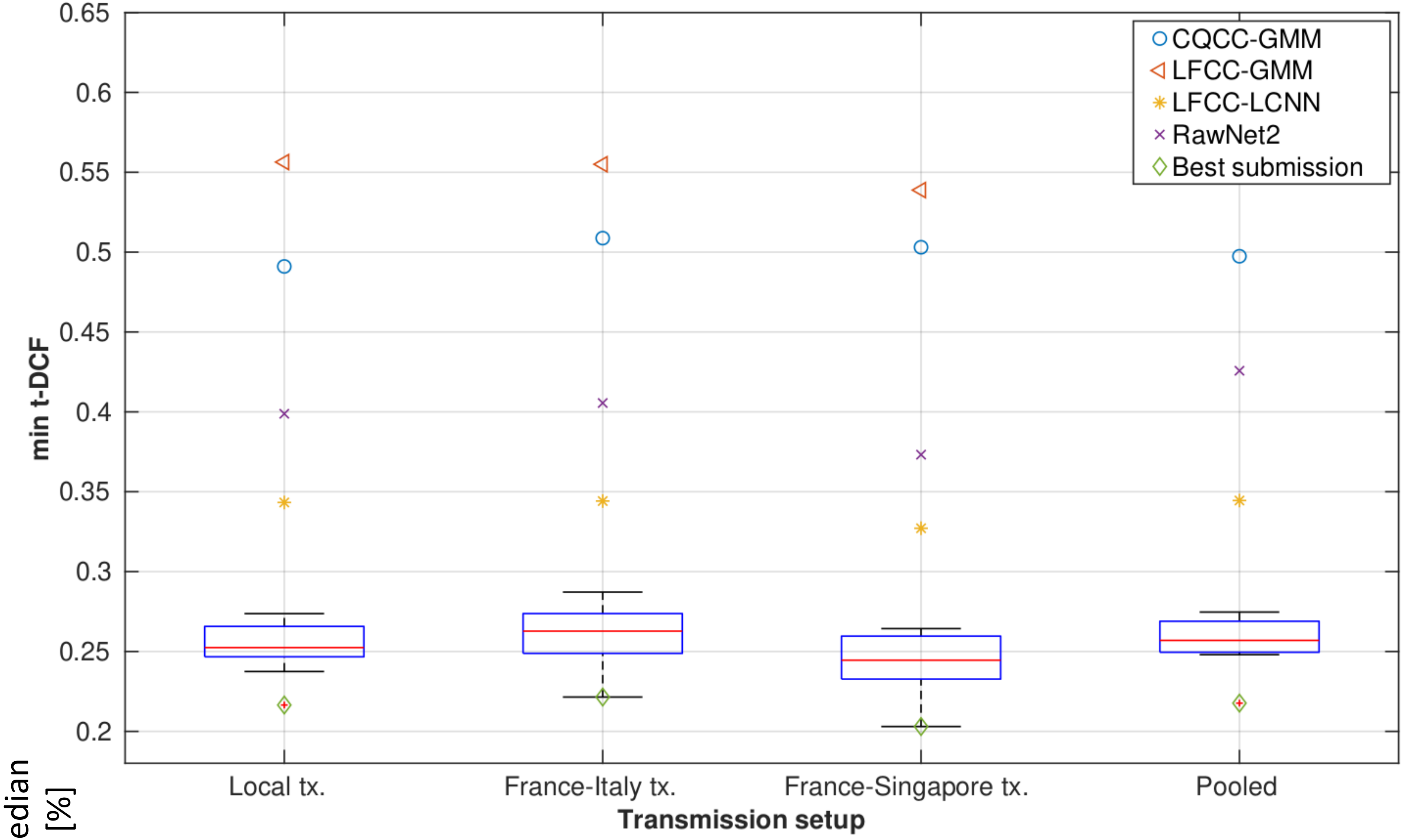
(top-10 submissions)



CM median EER [%]	None	OPUS	G.722	a-law	u-law	GSM	PSTN	Pooled
	1.87	2.31	2.14	2.33	2.32	3.48	5.59	3.17

LA results: transmission setup

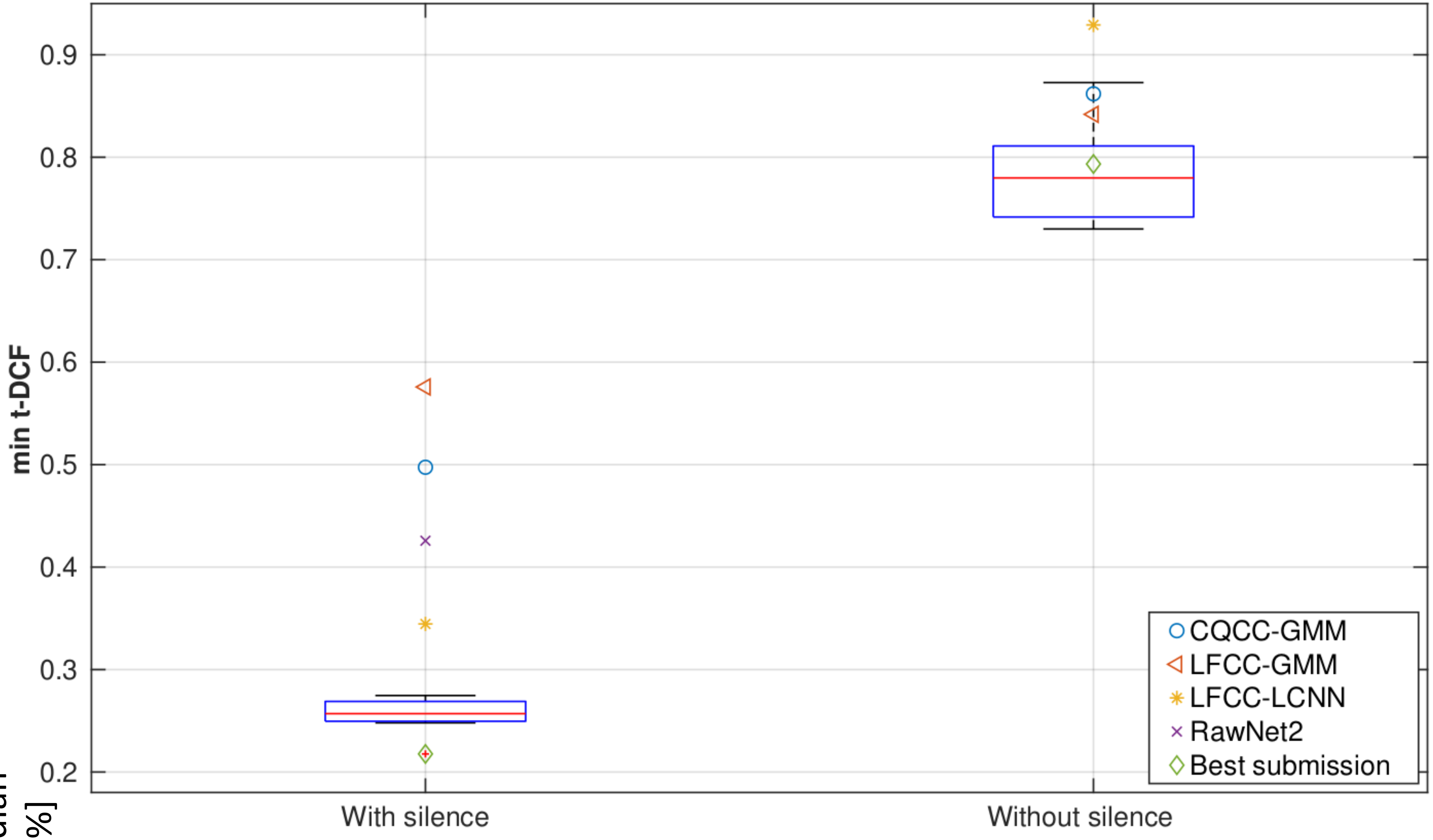
(top-10 submissions)



CM median EER [%]	Local tx.	France-Italy tx.	France-Singapore tx.	Pooled
	2.98	**3.51**	**3.22**	**3.17**

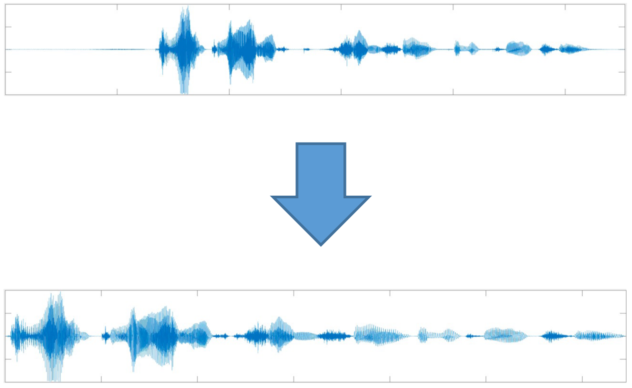
LA results: hidden track

(top-10 submissions)



CM median EER [%]	**3.17**	**27.03**

- **Silence removal from speech utterances (LA, PA and DF) [3-6]**



[3] B. Chettri et al. "A deeper look at Gaussian mixture model based anti-spoofing systems". In Proc. *IEEE ICASSP, 2018*.
[4] B. Chettri et al., "Ensemble models for spoofing detection in automatic speaker verification". In Proc. *Interspeech*, 2019.
[5] Y. Zhang et al., "The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System". In Proc. *Interspeech*, 2021.
[6] N. Müller et al. "Speech is silver, silence is golden: what do ASVspoof-trained models really learn?". In Proc. *ASVspoof 2021 Workshop*.

Physical access (PA)

Physical access

- Can CMs trained on simulated data (ASVspoof2019 PA) detect real spoofing data?

- We built the new evaluation set
  - real bona fide & replayed data recorded in physical rooms
  - ✓ room reverberation
  - ✓ background noise

- Compared with ASVspoof2019 PA evaluation set
  - same set of speakers and seed trials
  - controlled setup by design
  - ! real recording and replaying

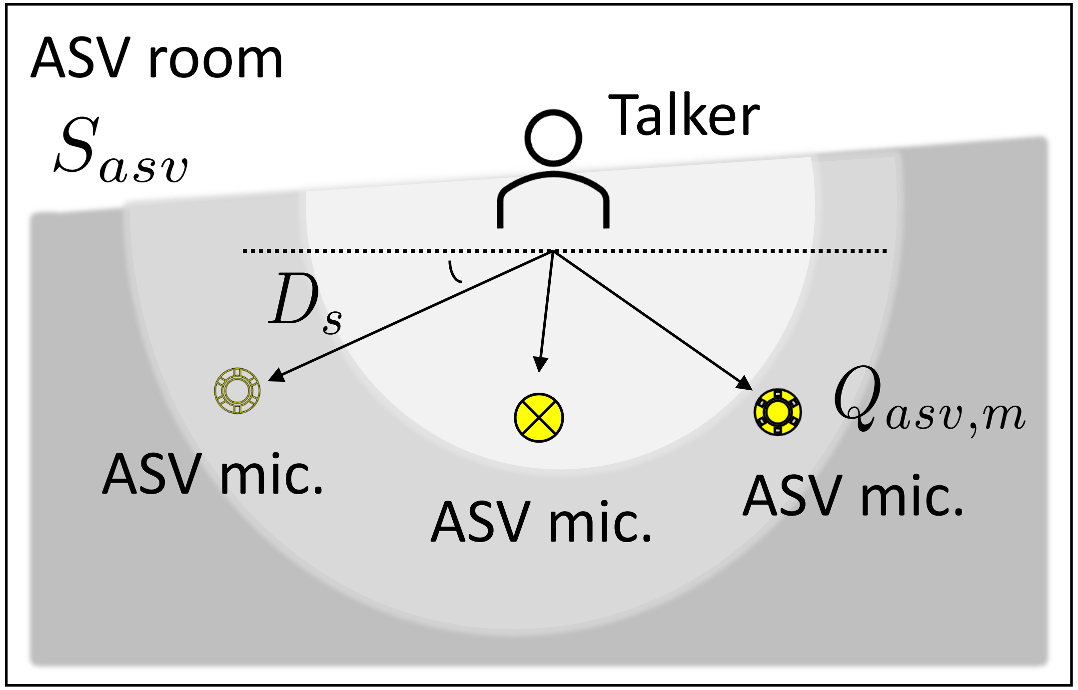
PA – Data collection

- Environmental factors:
  - ASV rooms S_{asv}
  - ASV mic. $Q_{asv,m}$
  - Talker-to-ASV-mic. distance D_s

	Cond.	Room size $w \times d \times h(m)$
S_{asv}	R1	$8.0 \times 8.0 \times 2.4$
	R2	$6.0 \times 5.0 \times 2.3$
	R3	$6.6 \times 5.0 \times 2.4$
	R4	$7.5 \times 7.7 \times 2.6$
	R5	$7.2 \times 4.0 \times 2.3$
	R6	$4.5 \times 6.5 \times 2.5$
	R7	$4.5 \times 2.4 \times 2.4$
	R8	$7.1 \times 4.8 \times 2.5$
	R9	$5.9 \times 4.0 \times 2.8$

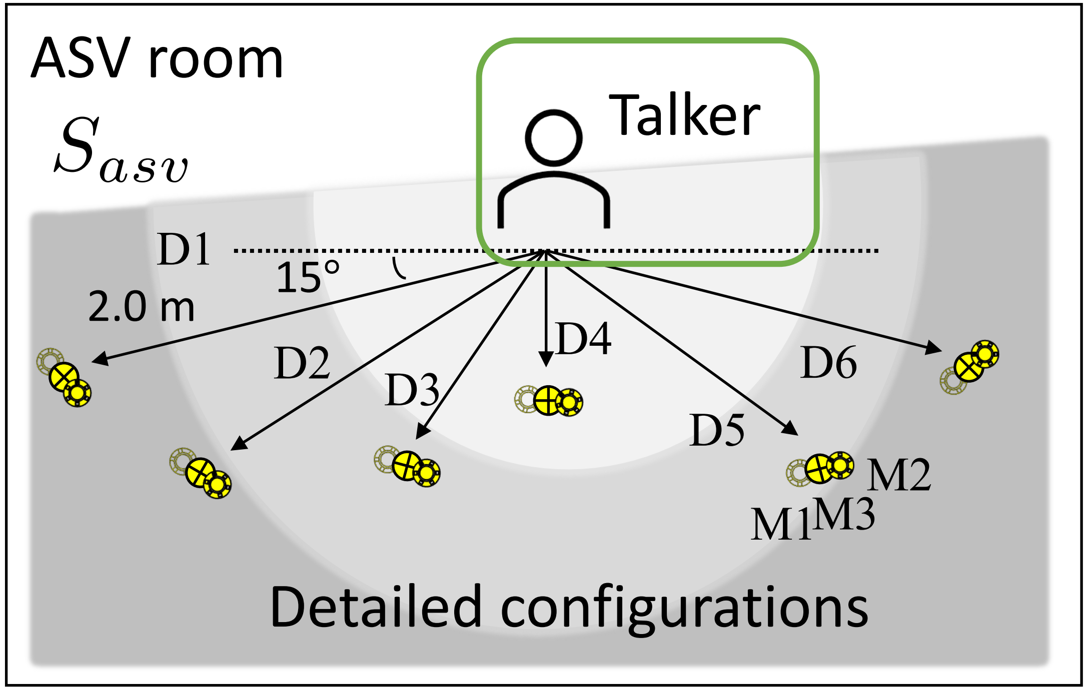
	Cond.	Device type
$Q_{\mathrm{asv,m}}$	M1	MPM-1000 (condenser)
	M2	Uber Mic (condenser)
	M3	iPad Air (MEMS)

	Cond.	Angle, Dis.(m)
D_s	D1	15°, 2.0
	D2	45°, 1.5
	D3	75°, 1.0
	D4	90°, 0.5
	D5	120°, 1.25
	D6	150°, 1.75



PA – Data collection

High-quality flat-response low-distortion loudspeaker for playing bona fide data



PA – Data collection

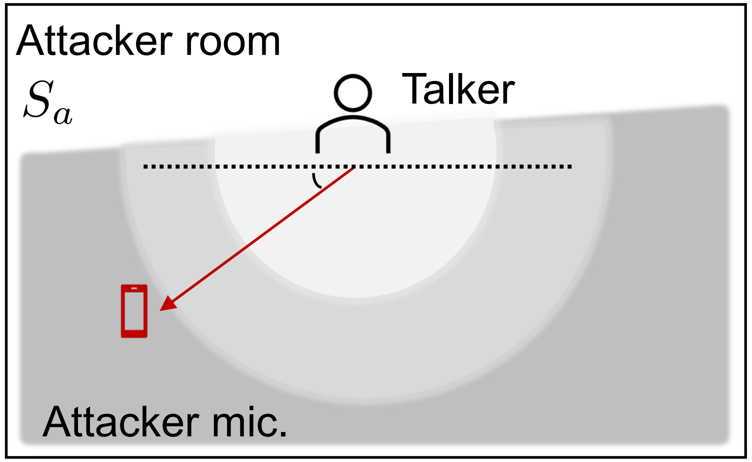
- Attacker factors (released)
  - Attacker room S_a
  - Replay-device-to-ASV-mic. distance ${D_s}'$

	Cond.	Room size $w \times d \times h(m)$
S_{asv} / S_{a}	R1 / r1	$8.0 \times 8.0 \times 2.4$
	R2 / r2	$6.0 \times 5.0 \times 2.3$
	R3 / r3	$6.6 \times 5.0 \times 2.4$
	R4 / r4	$7.5 \times 7.7 \times 2.6$
	R5 / r5	$7.2 \times 4.0 \times 2.3$
	R6 / r6	$4.5 \times 6.5 \times 2.5$
	R7 / r7	$4.5 \times 2.4 \times 2.4$
	R8 / r8	$7.1 \times 4.8 \times 2.5$
	R9 / r9	$5.9 \times 4.0 \times 2.8$

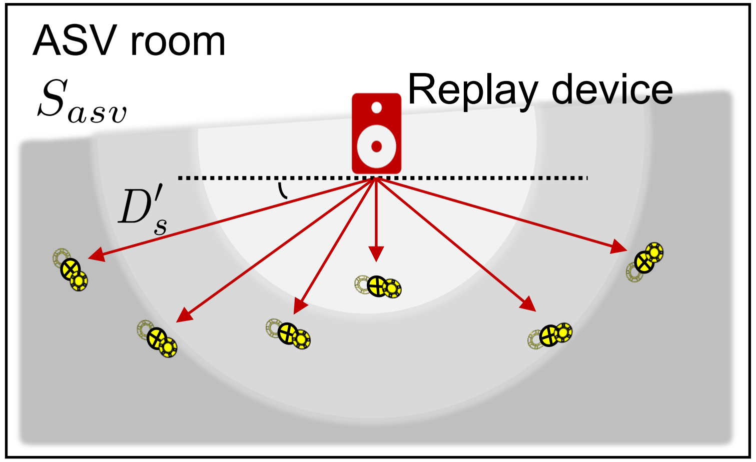
	Cond.	Device type
$Q_{\mathrm{asv,m}}$	M1	MPM-1000 (condenser)
	M2	Uber Mic (condenser)
	M3	iPad Air (MEMS)

	Cond.	Angle, Dis.(m)
D_s, D'_s	D1	15°, 2.0
	D2	45°, 1.5
	D3	75°, 1.0
	D4	90°, 0.5
	D5	120°, 1.25
	D6	150°, 1.75

Replay acquisition



Replay presentation



PA – Data collection

- Attacker factors (~~TBA~~)
  - Attacker microphone $Q_{a,m}$
  - Attacker-to-talker distance D_a
  - Replay device $Q_{a,s}$

	Cond.	Room size $w \times d \times h(m)$
	R1 / r1	$8.0 \times 8.0 \times 2.4$
	R2 / r2	$6.0 \times 5.0 \times 2.3$

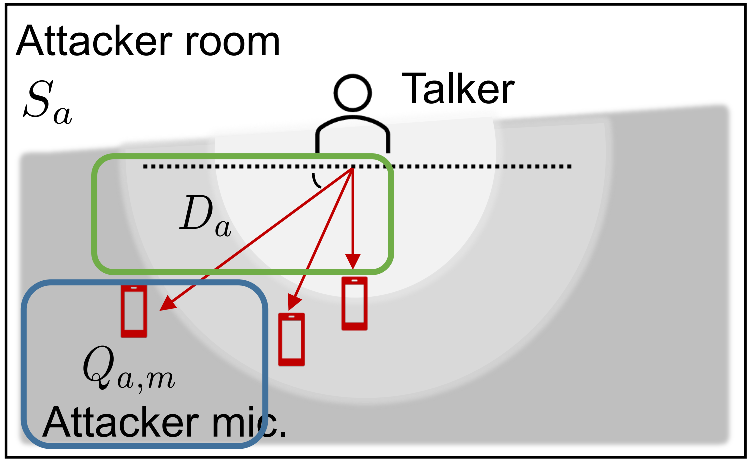
	Cond.	Device type
$Q_{\mathrm{asv,m}}$ / $Q_{\mathrm{a,m}}$	M1 / m1	MPM-1000 (condenser)
	M2 / m1	Uber Mic (condenser)
	M3 / m1	iPad Air (MEMS)

	Cond.	Angle, Dis.(m)
D_s, D'_s	D1	15°, 2.0
	D2	45°, 1.5
	D3	75°, 1.0
	D4	90°, 0.5
	D5	120°, 1.25
	D6	150°, 1.75

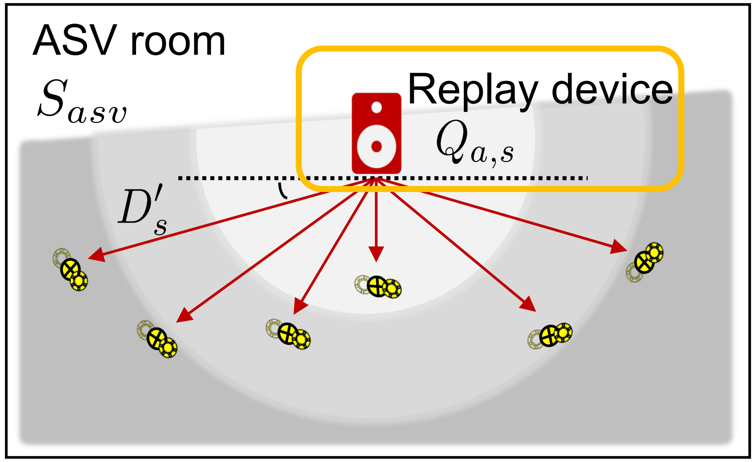
	Cond.	Device type
$Q_{a,s}$	s2	SONY
	s3	NEUMANN
	s4	GENELEC

	Cond.	Angle, Dis.(m)
D_a	d2	45°, 1.5
	d3	75°, 1.0
	d4	90°, 0.5

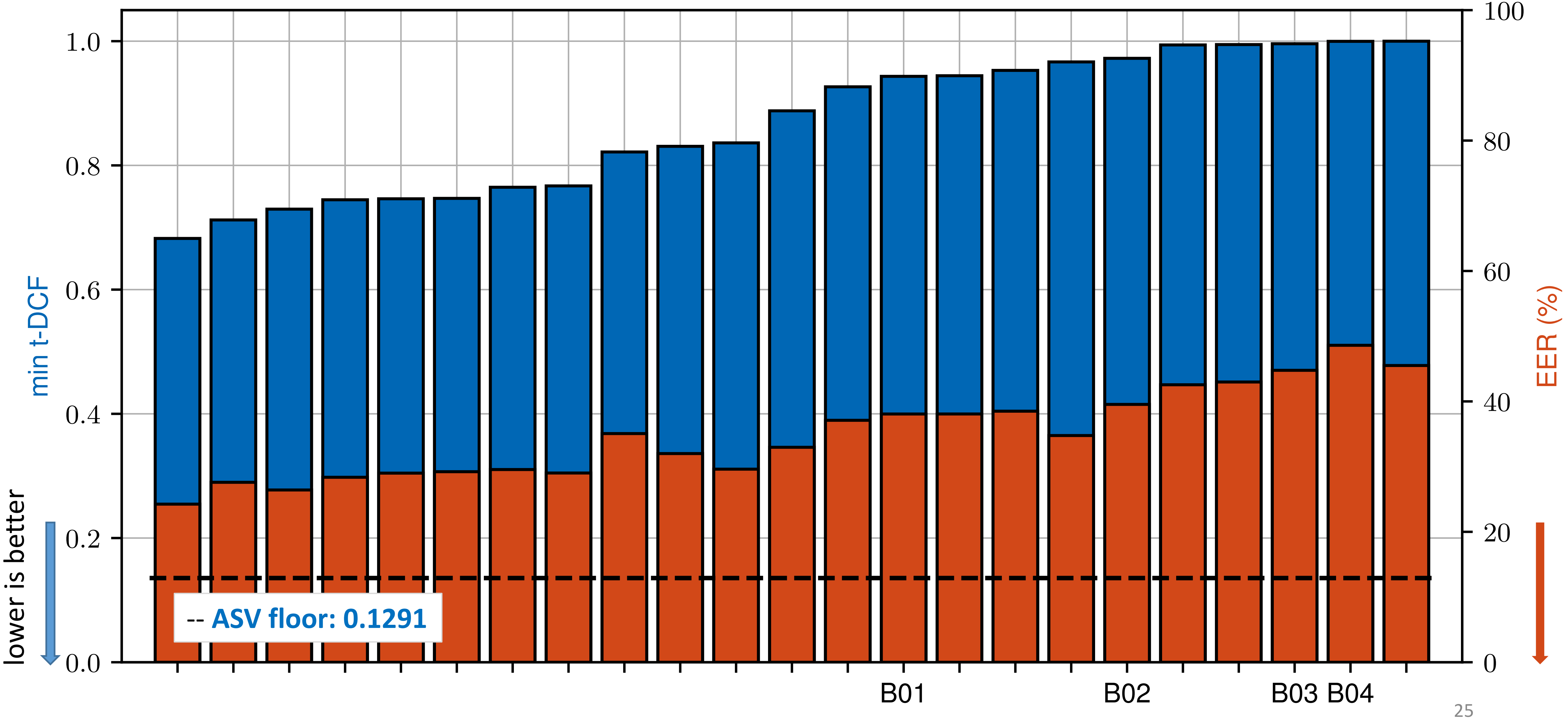
Replay acquisition



Replay presentation

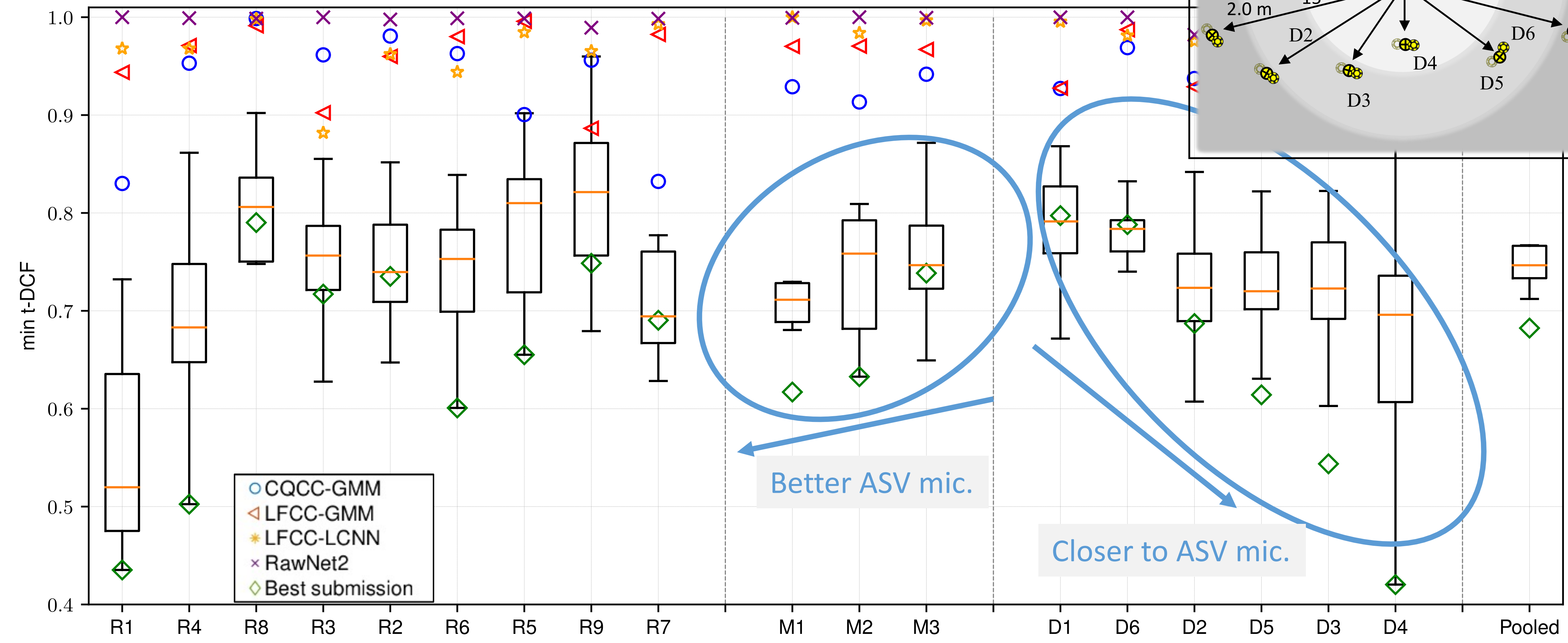


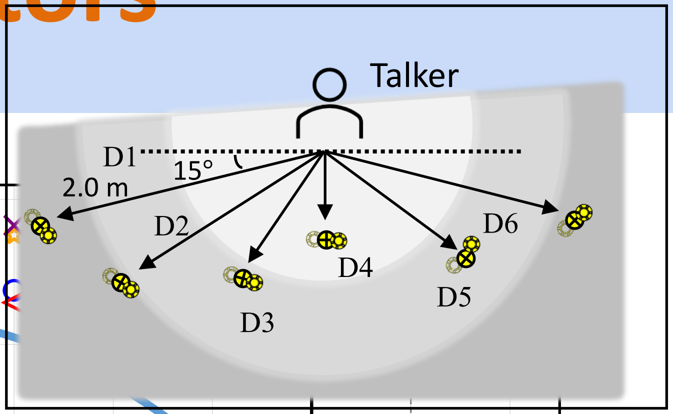
PA results: ranking



PA results: environmental factors

(top-10 submissions)





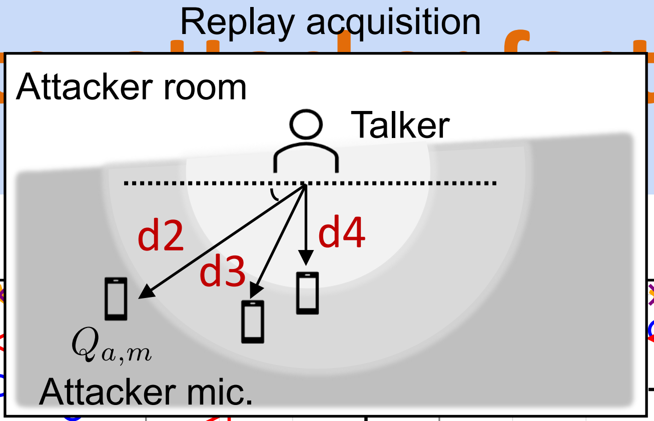
CM EER [%]	R1	R4	R8	R3	R2	R6	R5	R9	R7		M1	M2	M3		D1	D6	D2	D5	D3	D4		Pooled
	19.77	25.10	29.74	28.81	28.89	28.00	31.25	32.66	27.03		27.68	29.25	28.47		30.42	30.97	27.07	27.82	28.84	25.93		29.01

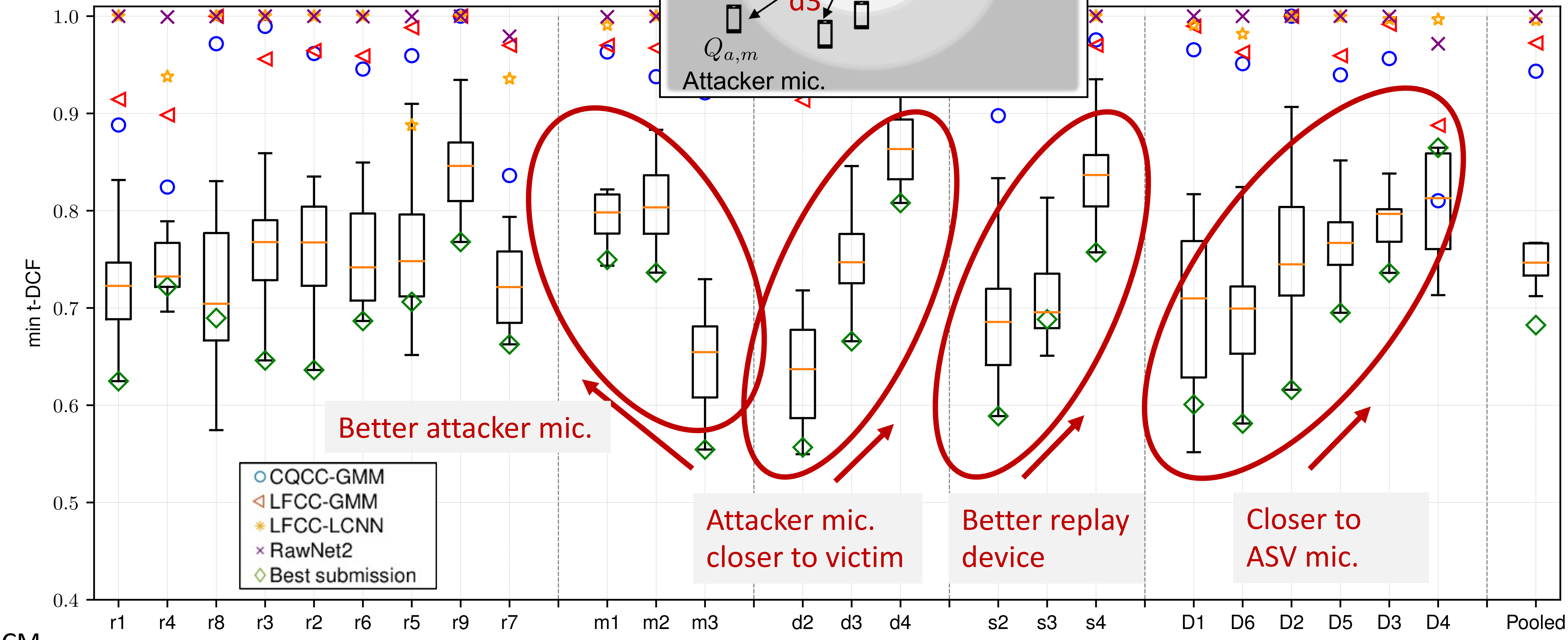
Larger ASV room size

	Cond.	Device type
$Q_{\text{asv,m}}$	M1	MPM-1000 (condenser)
	M2	Uber Mic (condenser)
	M3	iPad Air (MEMS)

Closer to ASV mic.

PA result… …tors





	r1	r4	r8	r3	r2	r6	r5	r9	r7	m1	m2	m3	d2	d3	d4	s2	s3	s4	D1	D6	D2	D5	D3	D4	Pooled
CM EER [%]	27.52	27.6	25.92	30.69	29.16	28.26	27.34	34.43	27.94	30.81	31.2	23.86	22.75	27.77	34.44	25.35	26.04	33.42	26.93	25.84	28.13	29.52	30.77	31.84	29.01

Larger attacker room size

	Cond.	Device type
$Q_{a,m}$	m1	MPM-1000 (condenser)
	m2	Uber Mic (condenser)
	m3	iPad Air (MEMS)

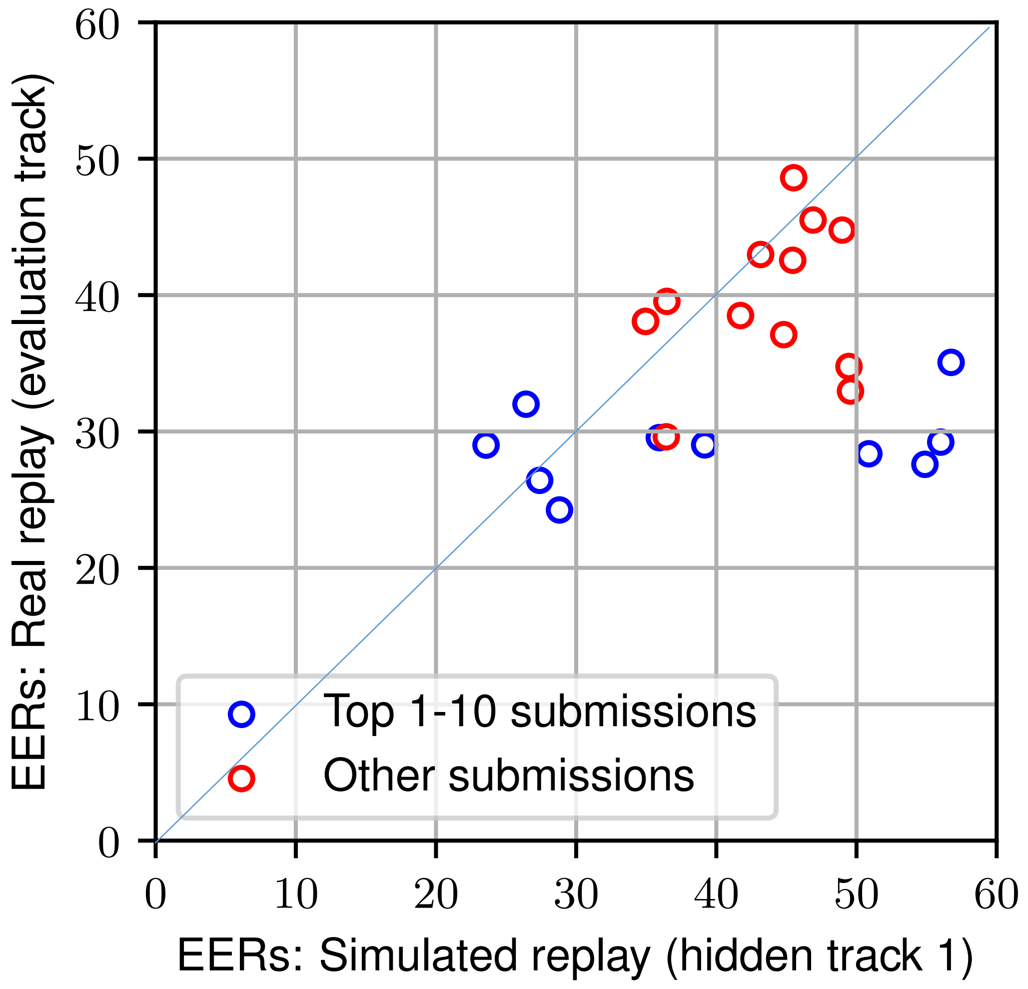
	Cond.	Angle, Dis.(m)
D_a	d2	45°, 1.5
	d3	75°, 1.0
	d4	90°, 0.5

	Cond.	Device type
$Q_{a,s}$	s2	SONY
	s3	NEUMANN
	s4	GENELEC

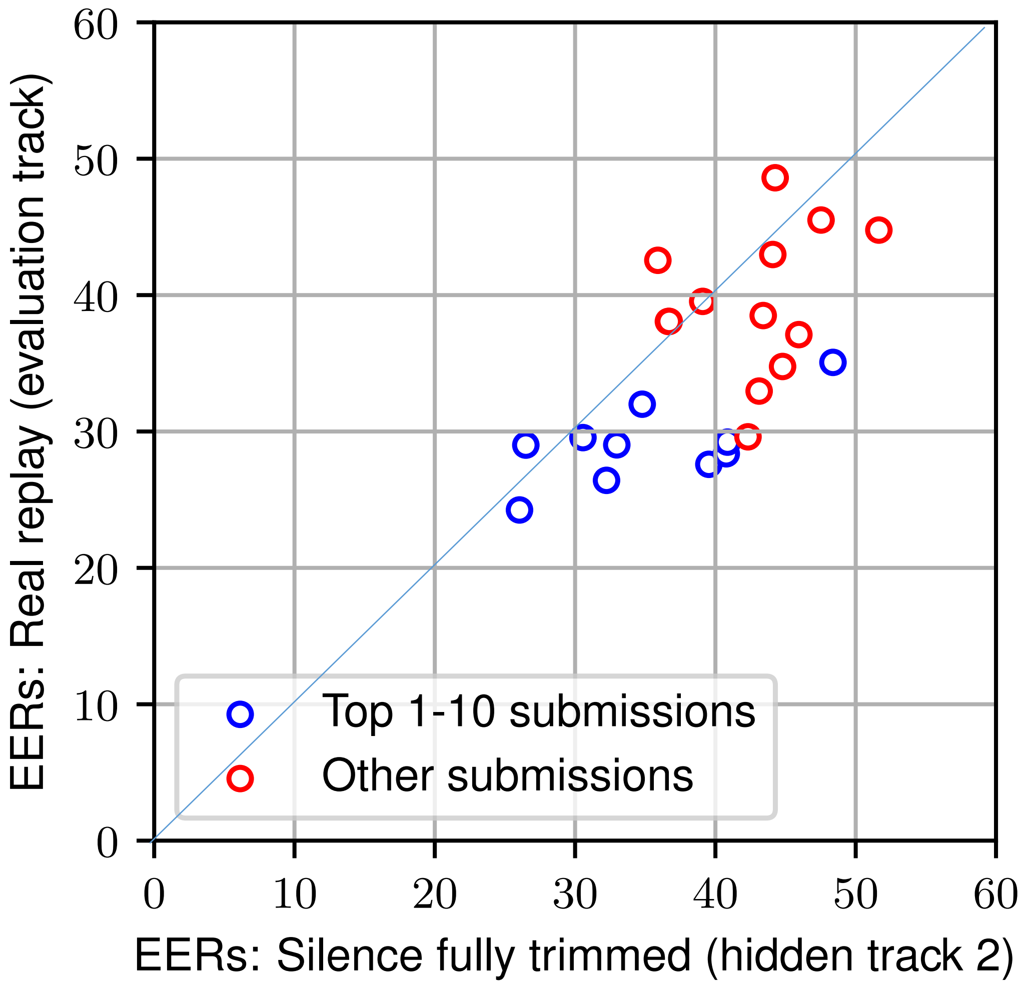
Closer to ASV mic.

PA results: hidden track

- Hidden track 1: simulated PA data

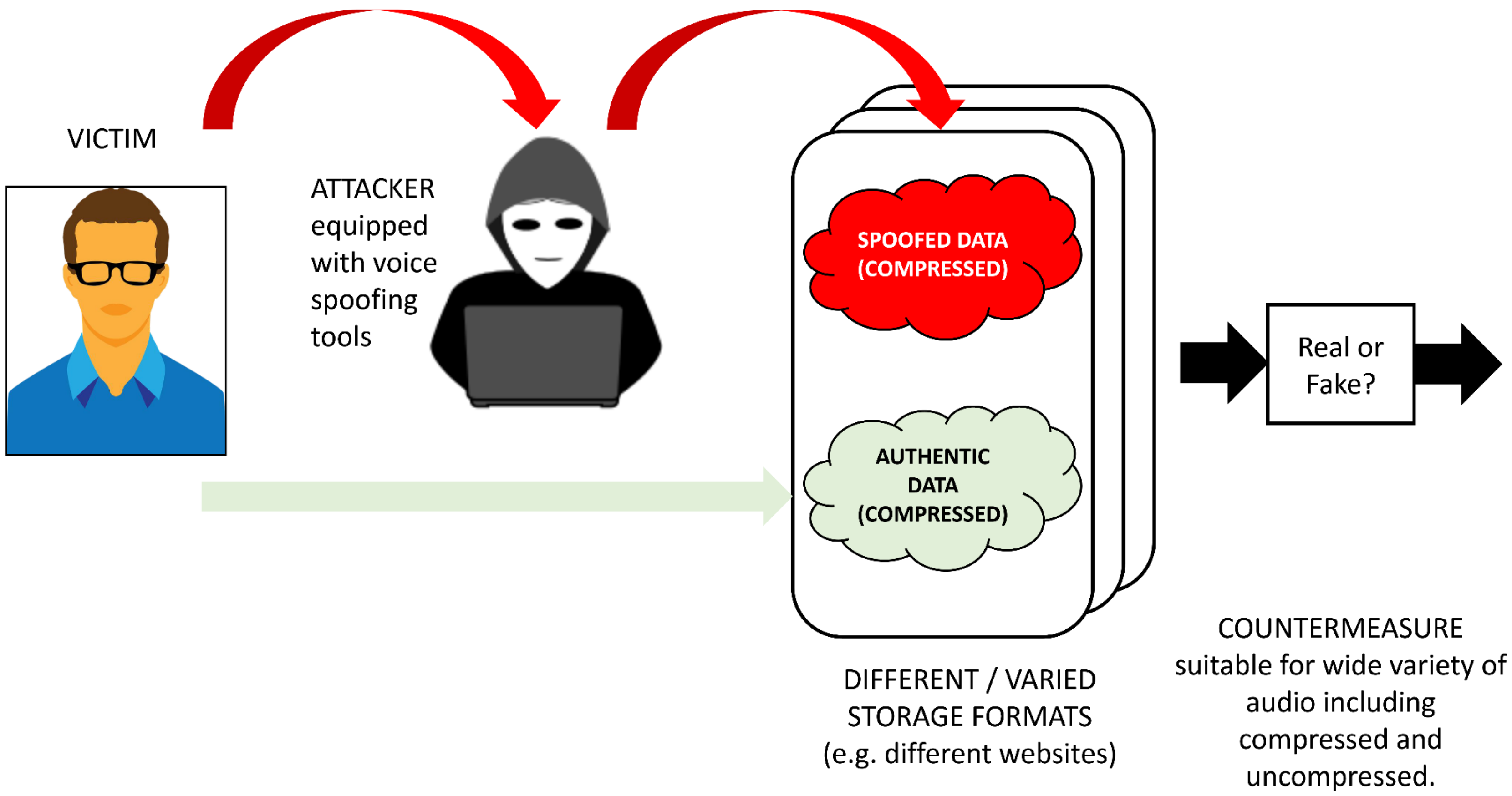


- Hidden track 2: without silence



Speech deepfake (DF)

Speech deepfake



- based on audio-data processed with different lossy compression schemes typically used for media storage
- collected from several sources covering a large number of attacks.

DF - Conditions

Compression Methods:

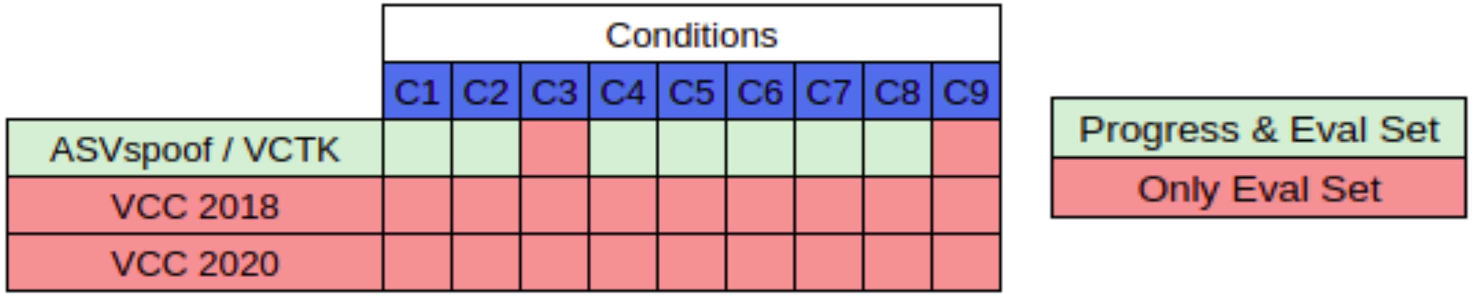
Cond.	Compression (Quality)	VBR settings (kbps)
C1	None	-
C2	mp3 (low)	~80-120
C3	mp3 (high)	~220-260
C4	m4a (low)	~20-32
C5	m4a (high)	~96-112
C6	ogg (low)	~80-96
C7	ogg (high)	~256-320
C8	mp3 (low) => m4a (high)	~80-120=>~96-112
C9	ogg (low) => m4a (high)	~80-96=>~96-112

No compression (C1)

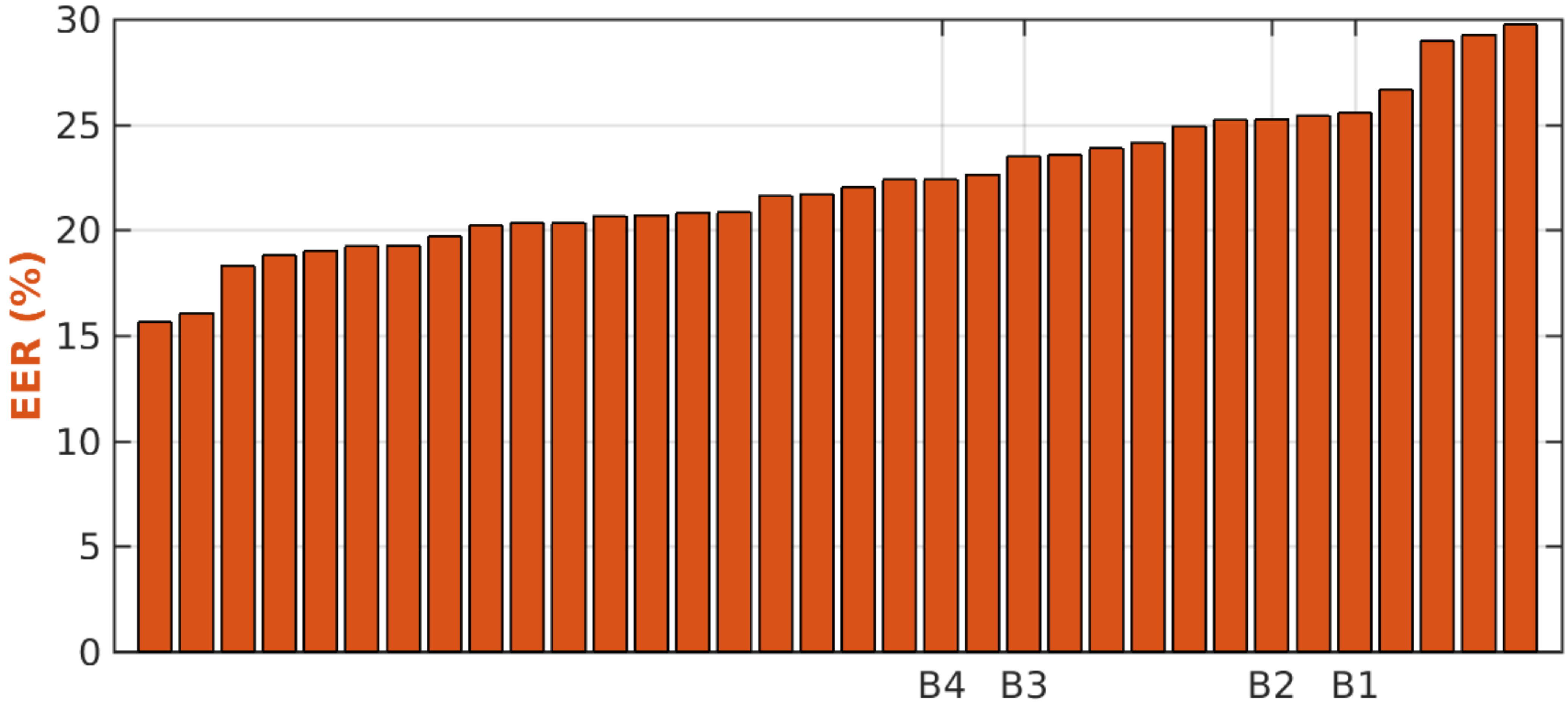
Single compression (C2–C7)

Double compression (C8–C9)

Audio data:

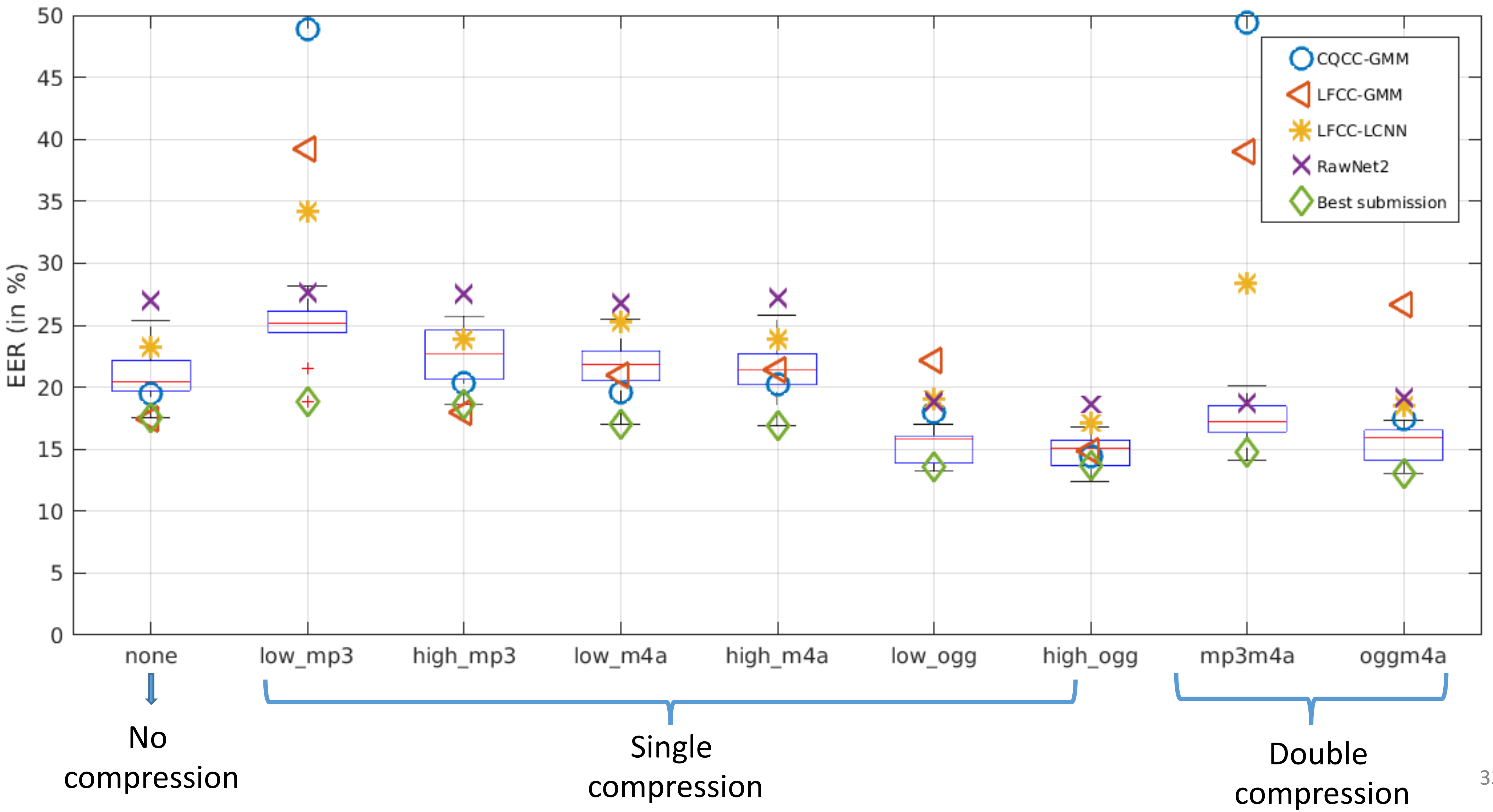


DF results: ranking



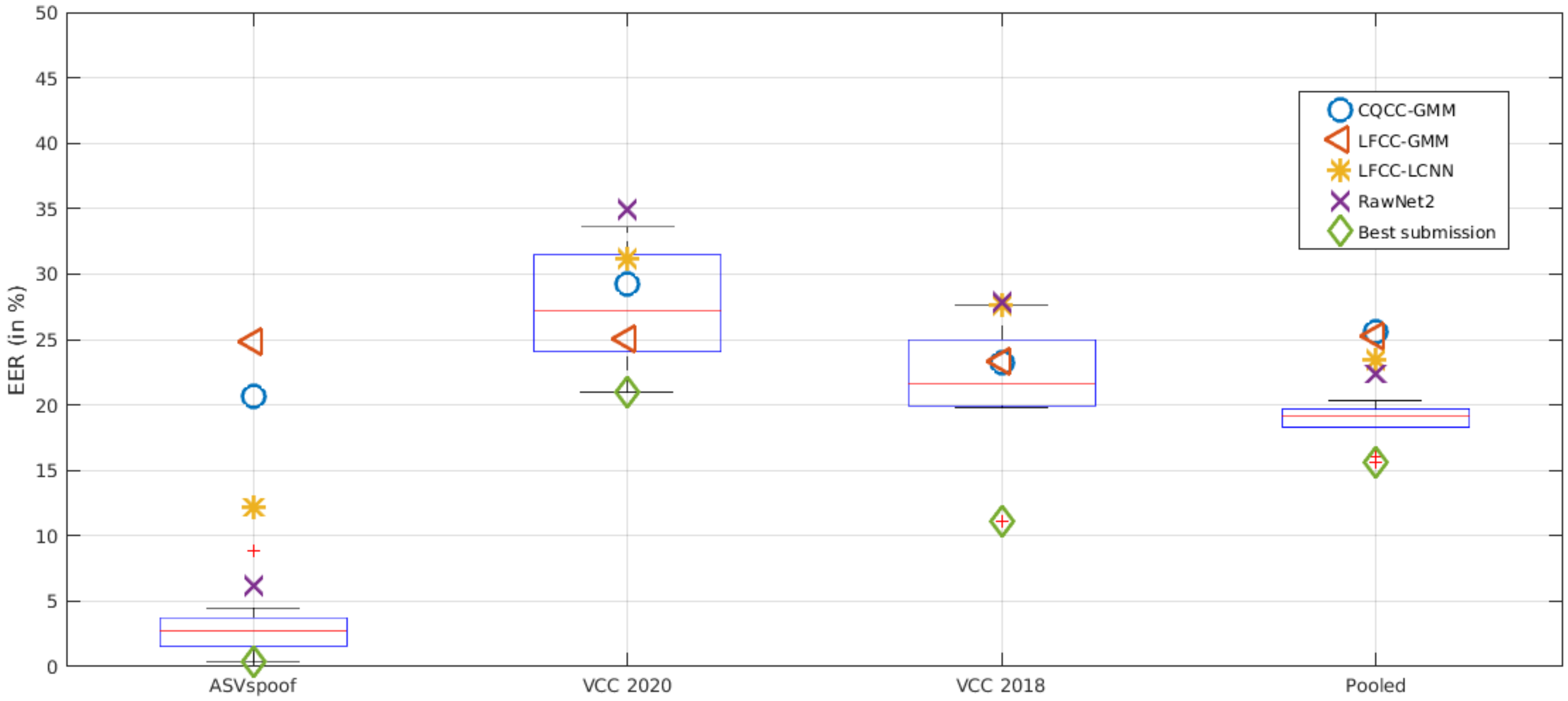
DF results: compression conditions

(top-10 submissions)



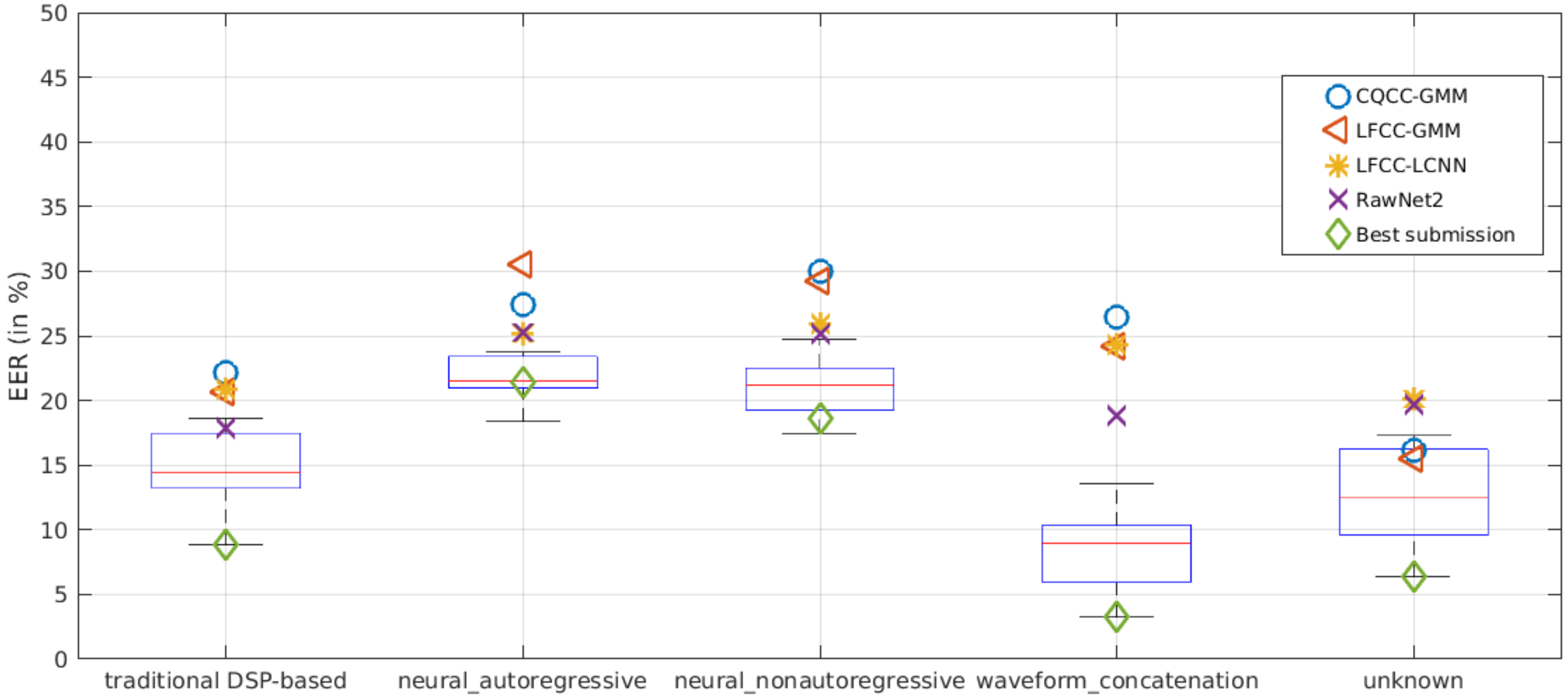
DF results: source dataset

(top-10 submissions)



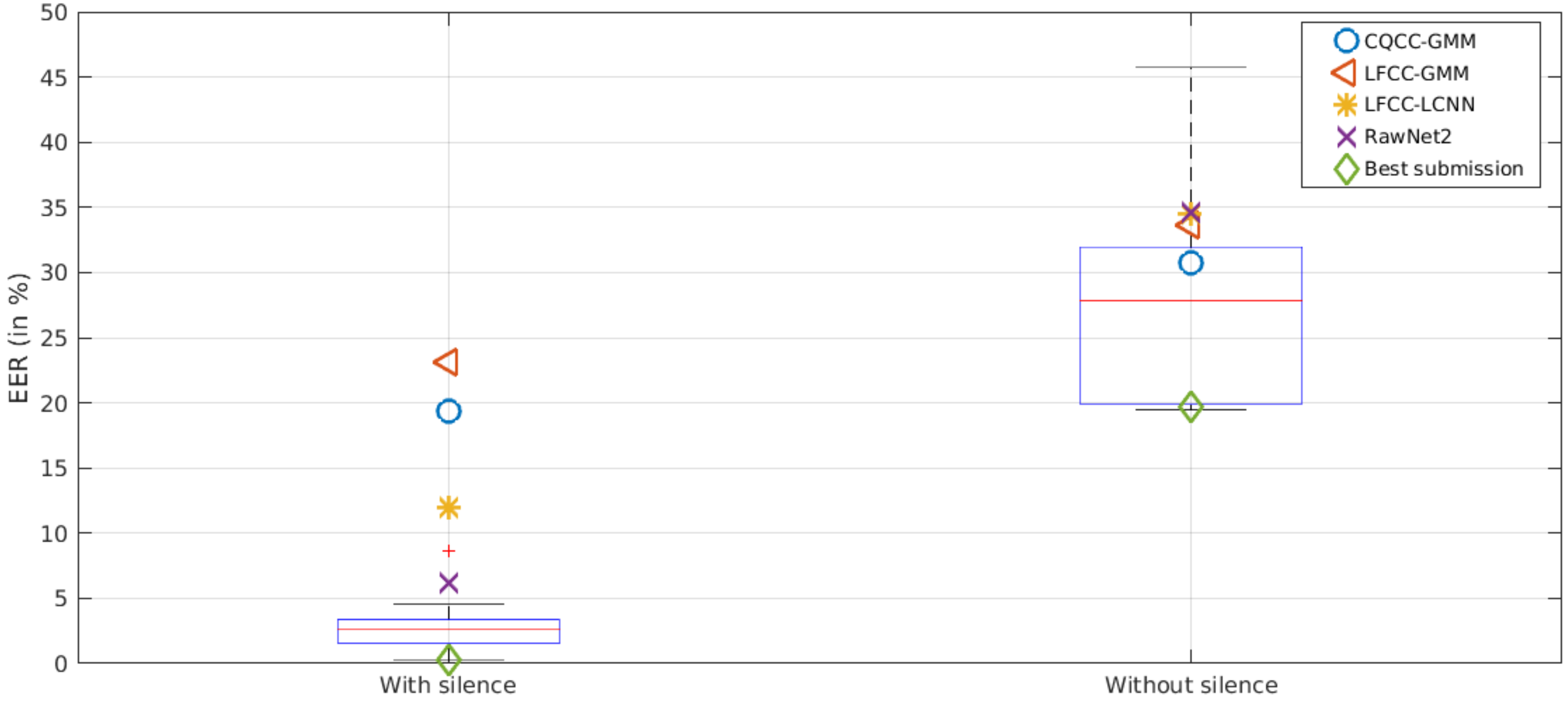
DF results: vocoder type

(top-10 submissions)



DF results: hidden track

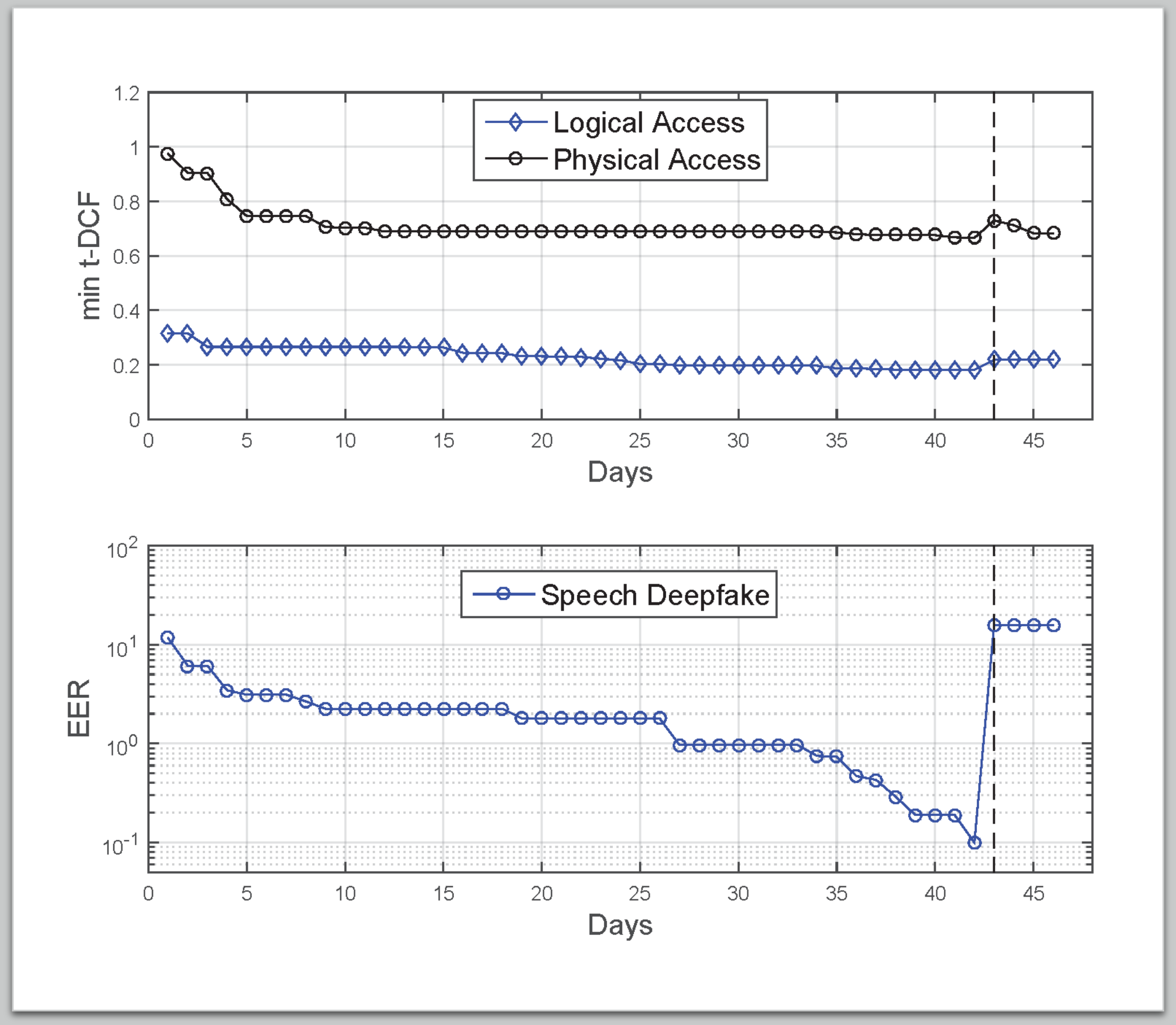
(top-10 submissions)



Conclusions

Overall results

- On the progress set, the reduction in min t-DCF amounts to 42% and 32% on the LA and PA tasks.
- On the DF task, EER was reduced from 11.6% to 0.10% during the progress phase, which proven to be illusive when the EER shoot up to 15.6% in the evaluation phase.



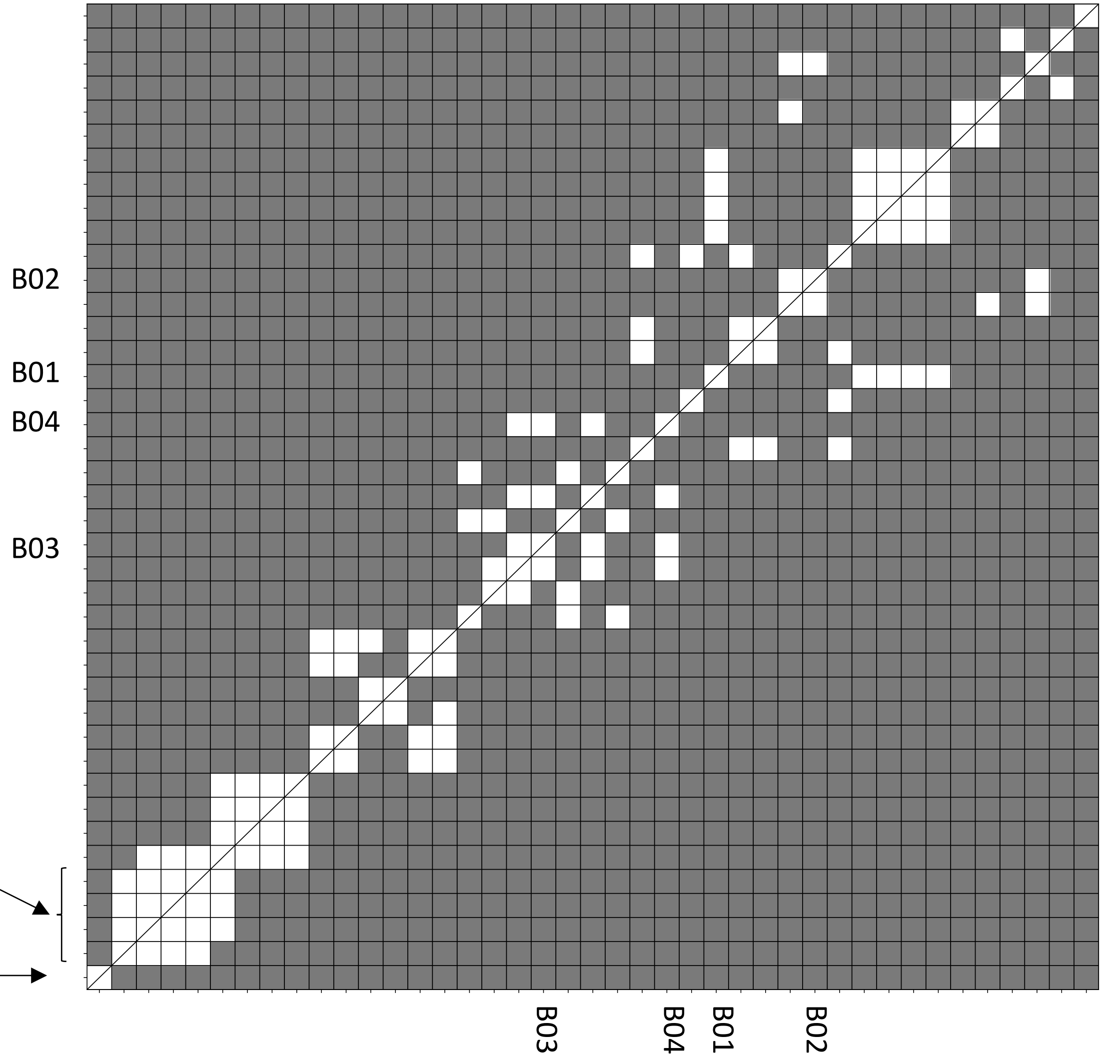
Statistical significance

- Pairwise significant test
  - On the basis of EERs [7]
  - Holm-Bonferroni correction applied
  - Grey block indicates statistically significant difference at $\alpha = 0.05$
  - Otherwise, block is in white

Top2-4 submissions are not statistically significantly different from each other

Top-1 submission is different from others

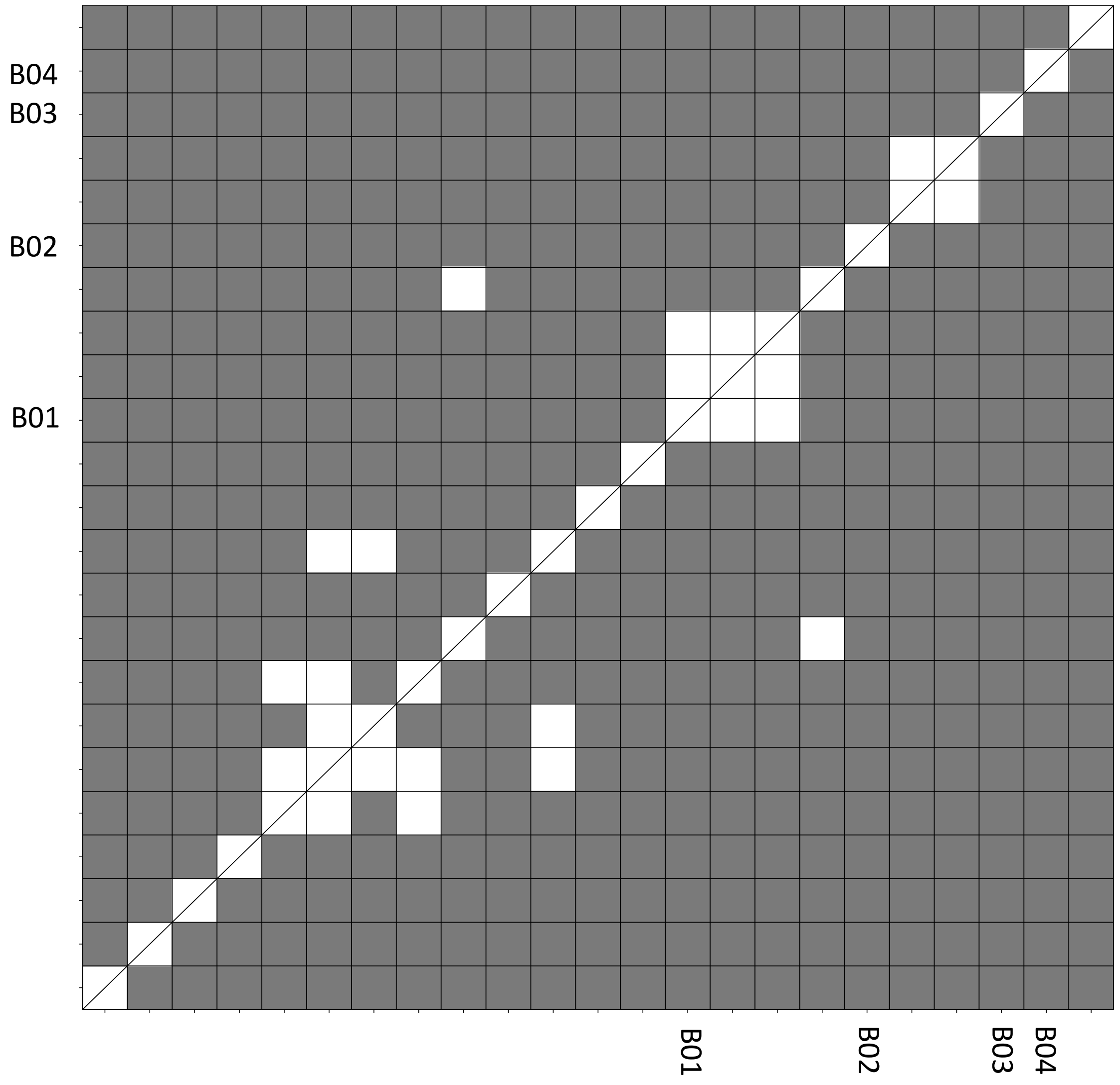
LA



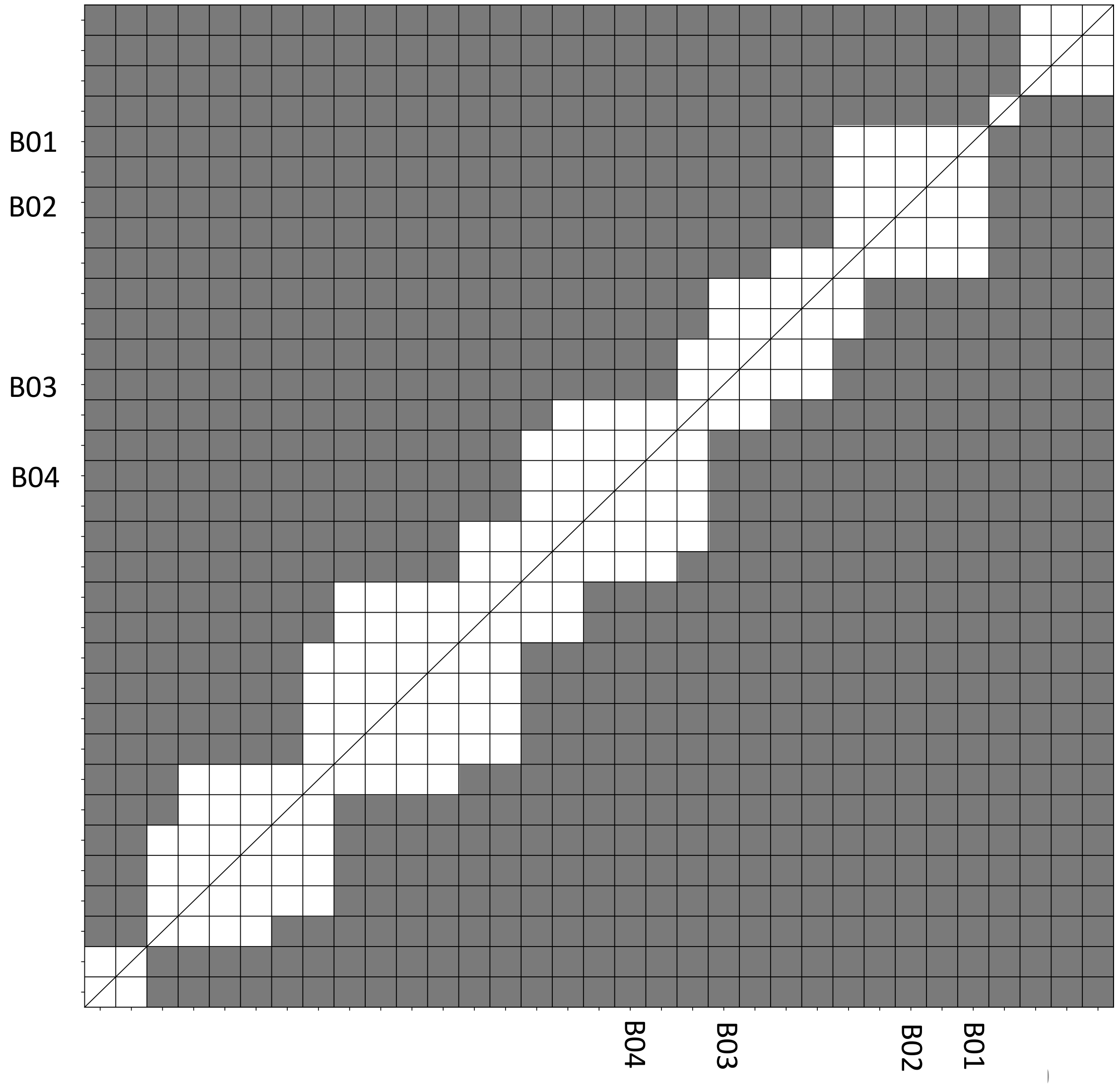
[7] Bengio, S. & Mariéthoz, J. "A statistical significance test for person authentication". in *Proc. Speaker Odyssey* 2004.

Statistical significance

PA



DF



Findings

- substantial improvements over baselines
- performance close to the ASV floor (LA)
  - despite new channel variability
  - despite absence of training/dev data
- challenges remain (esp. PA)
  - meta-data (e.g. room conditions) might be needed to improve performance
  - use of simulated data to learn classifiers capable of detecting real PA attacks
- some gaps between performance for progress and eval. set (esp. DF)
  - source dataset & compression quality are key factors
- successful adoption of new format, inc. CodaLab & leaderboard
- **encouraging results, despite increased difficulty**

Acknowledgement

Contribution to ASVspoof 2021 LA database (LA-C3 condition)

- We wish to thank:
  - contributors
  - authors/presenters
  - participants
  - reviewers

Contribution to ASVspoof 2019 LA database

Special thanks to Sébastien Le Maguer for preparing the workshop proceedings now available from the ISCA archive