

RESPECT

REliable, Secure and Privacy preserving multi-biometric pErson authentiCaTion



Presentation Attack Detection for Voice Biometrics: When Audio Meets Image

Massimiliano Todisco

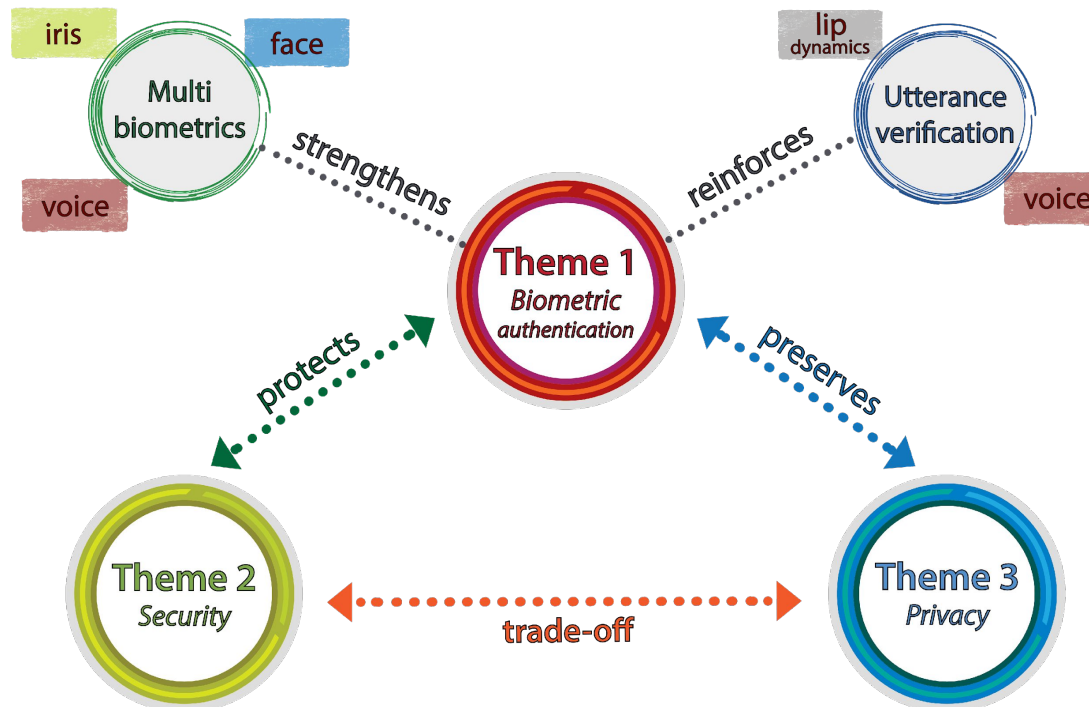
EURECOM, France

The RESPECT project



objectives

- **Theme 1**: multi-biometric authentication + utterance verification
- **Theme 2**: security: countermeasures against spoofing attacks
- **Theme 3**: privacy: protection of sensitive data



- **French-German collaborative project**
 - 4 universities and research centres
 - project start: 1st April 2019
 - duration: 36 months

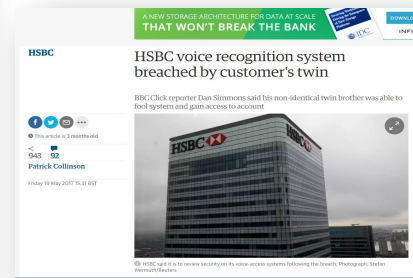


voice biometric spoofing/presentation attacks

[ISO/IEC 30107-1:2016]

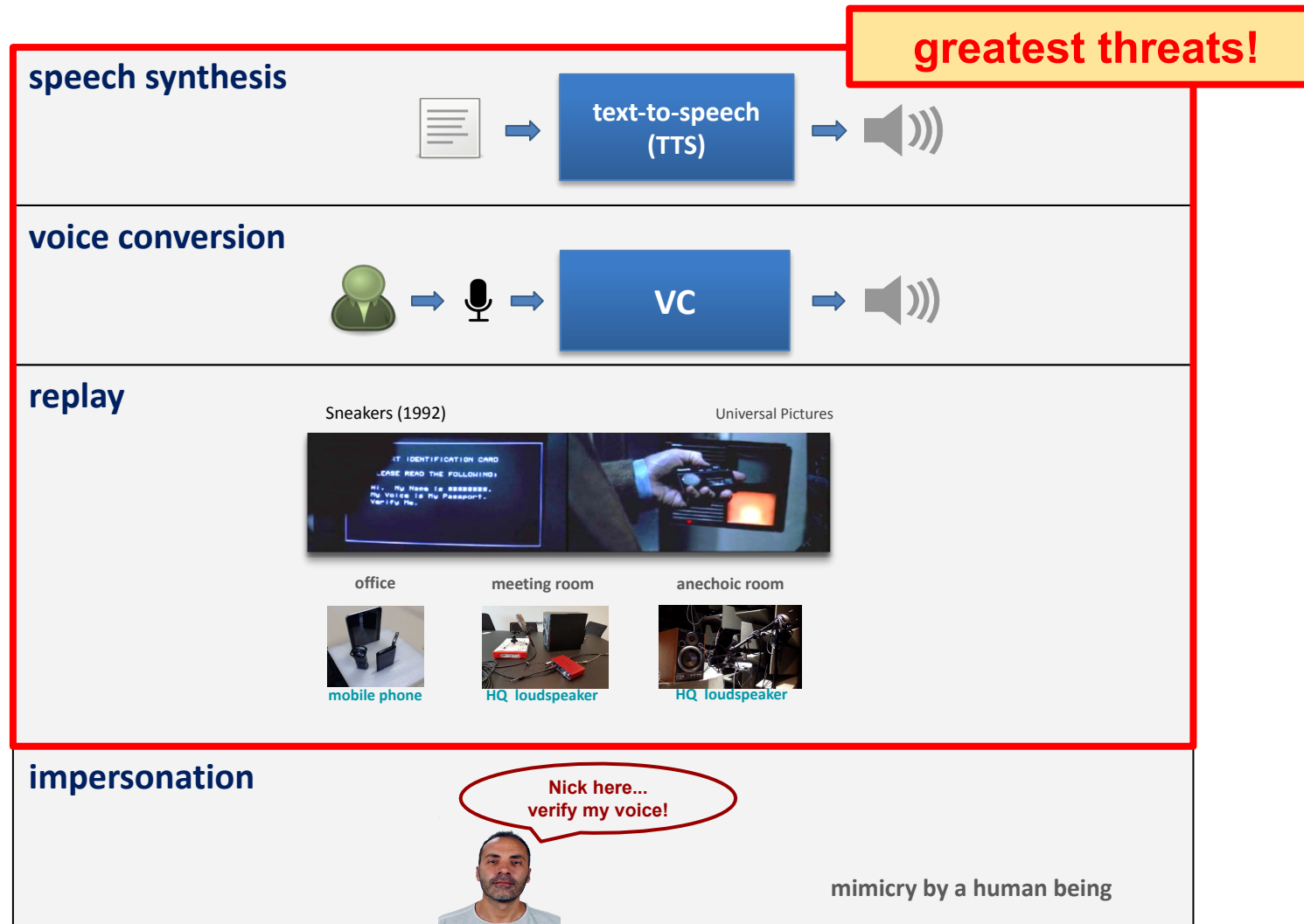


- security in voice biometrics is becoming a necessity



voice biometric spoofing/presentation attacks

[ISO/IEC 30107-1:2016]

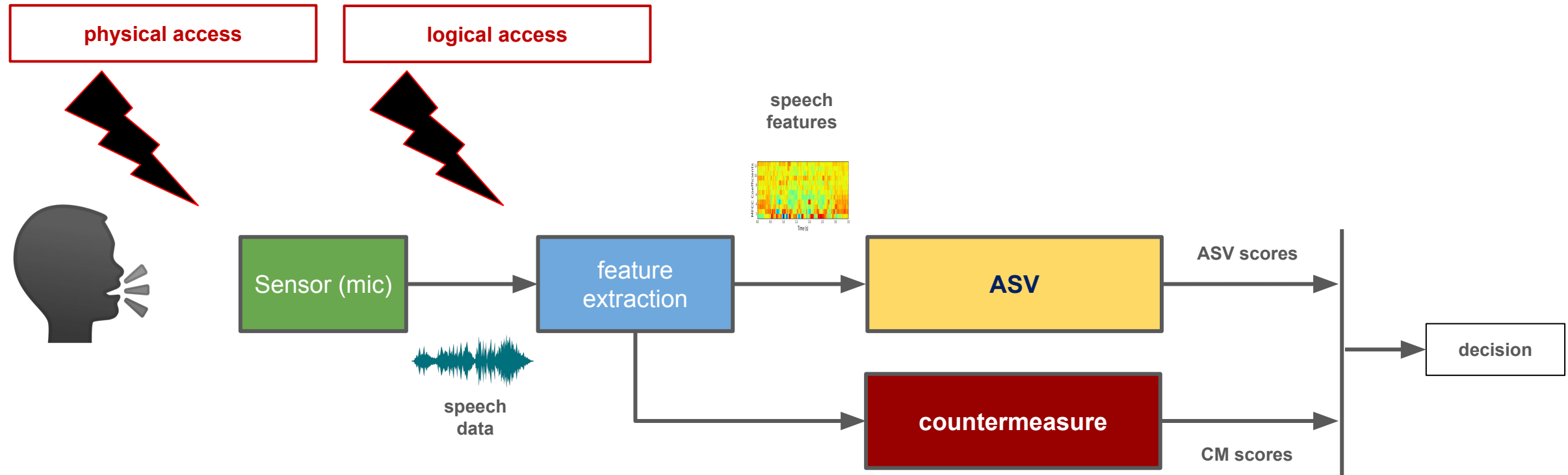


voice biometric spoofing/presentation attacks

[ISO/IEC 30107-1:2016]



- sensor level: before and/or after the microphone

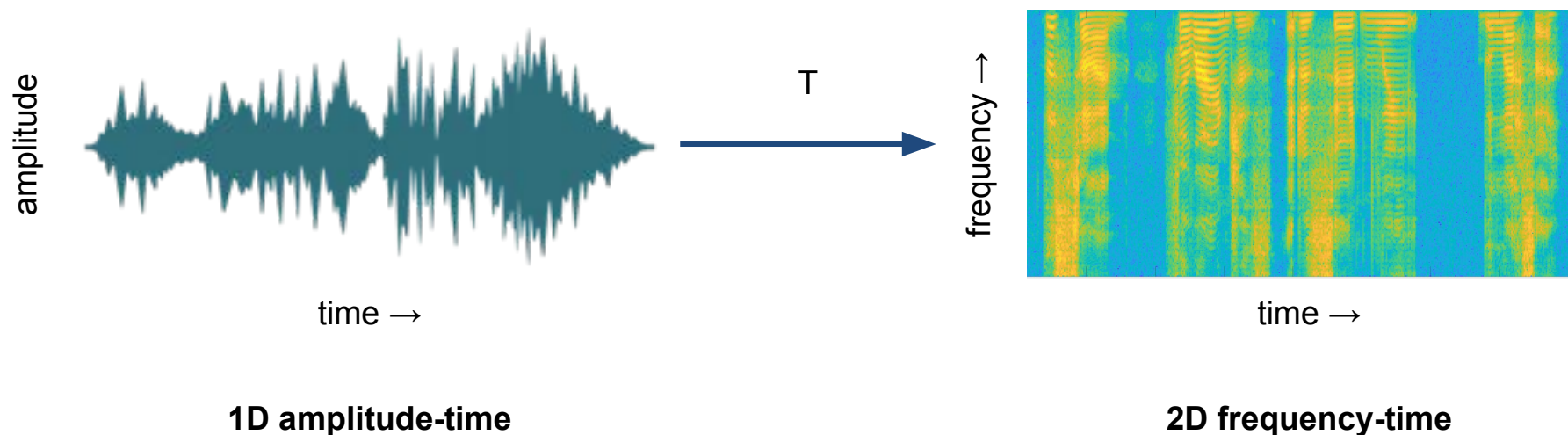


detecting artefacts in ASV spoofing attacks



motivation

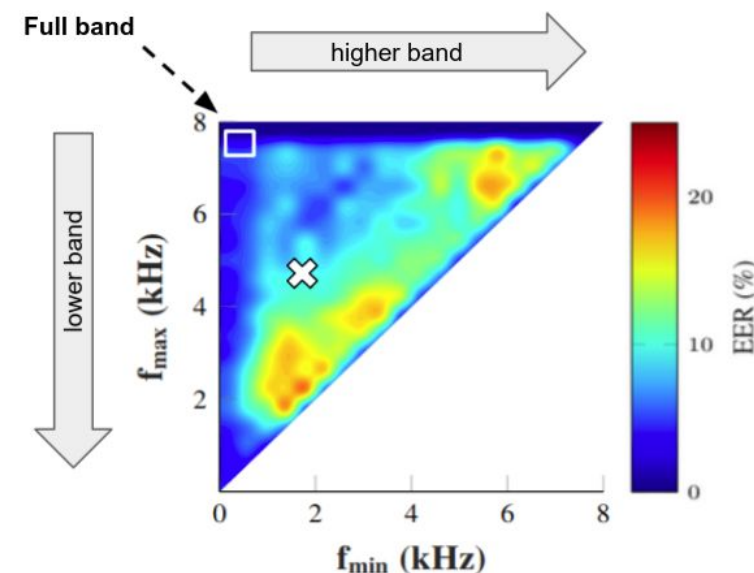
- spoofing artefacts can be localised in the
 - behaviour of amplitude samples
 - spectrum, e.g. high-band, mid-band or low-band
- conventional cepstral analysis smooths information across the full band and dilutes localised information
- seeking reliable detection with features that emphasise information at the sub-band level



detecting artefacts in ASV spoofing attacks



- **understanding constant-Q cepstral coefficients (CQCCs) [1]**
 - proposed heat-map visualisation methods to detect informative sub-bands
 - analysed impact of spectro-temporal resolution in spoofing detection
- **... and leveraging that knowledge [2]**
 - developed automatic sub-band detection selection methods
 - optimised filterbank density for near state-of-the art (SOTA) performance with a simple GMM-based system



heat-map based artefact visualisation based on band-pass filtering, where a working point, indicated by the cross, involves a cut-in and cut-off frequencies of filters.

1. Tak, H., Patino, J., Nautsch, A., Evans, N. and Todisco, M., An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification. In proc. Speaker Odyssey: The Speaker Recognition Workshop, 2020
2. Tak, H., Patino, J., Nautsch, A., Evans, N. and Todisco, M., Spoofing attack detection using the non-linear fusion of sub-band classifiers. In proc. Interspeech 2020

detecting artefacts in ASV spoofing attacks



- **further work explored texture analysis based approaches [1]**
 - from audio to image domain
 - Fourier and constant Q transformation and quantisation
- **... and leveraging established image processing methods**
 - fisher vector feature space based on a generative model
 - Binarized Statistical Image Features (BSIF)

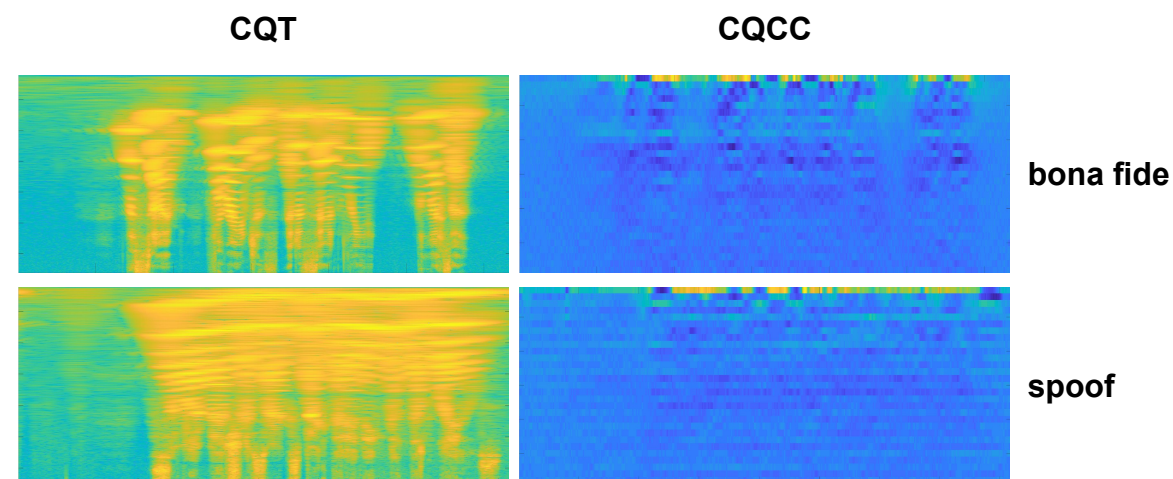


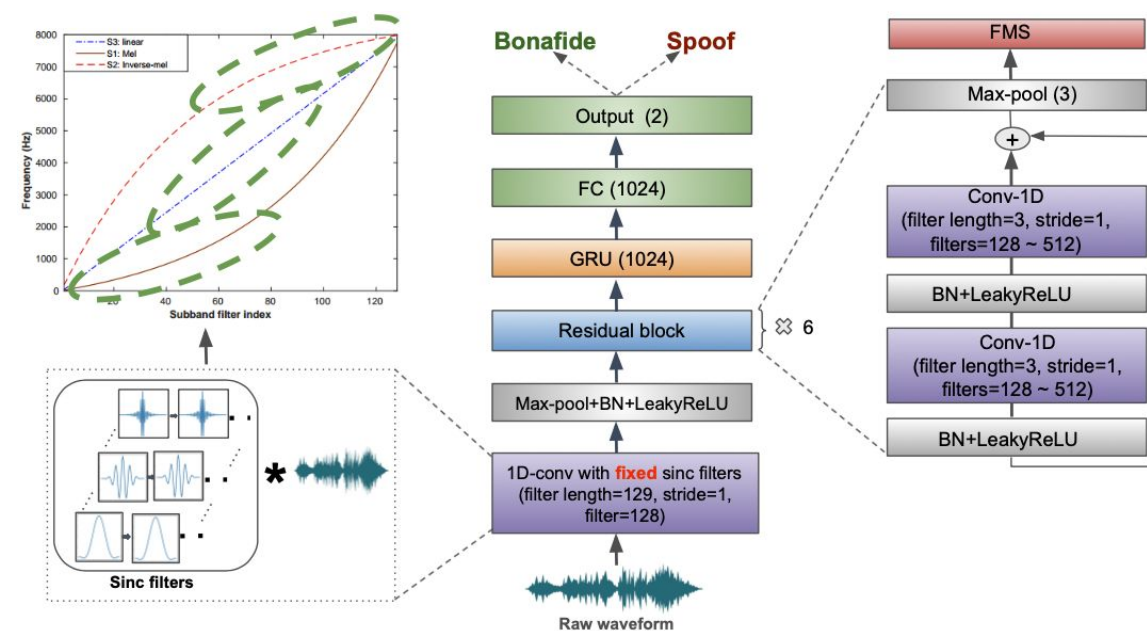
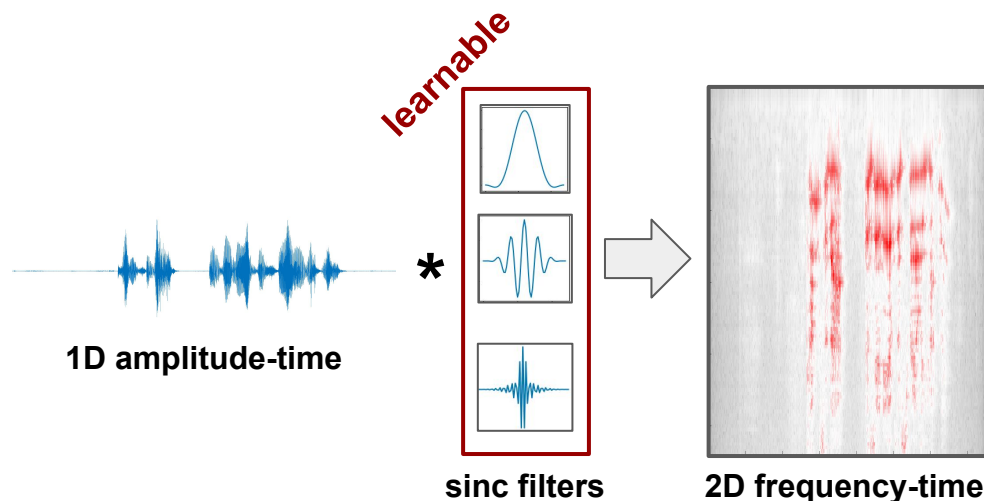
image features in [1]

1. Gonzalez-Soler L. J., Patino J., Gomez-Barrero M., Todisco M., Busch C., Evans N., 2020, December. Texture-based Presentation Attack Detection for Automatic Speaker Verification. In proc. IEEE International Workshop on Information Forensics and Security 2020.

detecting artefacts in ASV spoofing attacks



- **exploring end-to-end DNNs for spoofing detection [1]**
 - specific spoofing attacks would not be detected using traditional cepstral analysis
 - end-to-end, high-time resolution processing eased the way
- **... and learnable 1D → 2D speech transformation**
 - sinc filters enabled complementary systems close to SOTA



overview of contributions presented in [1]

ASVspoof initiative: databases and challenges



ASVspoof

Automatic Speaker Verification and
Spoofing Countermeasures Challenge



2015 - 16 participants



2017 - 49 participants



2019 - 98 participants

ASVspoof 2021
on-going

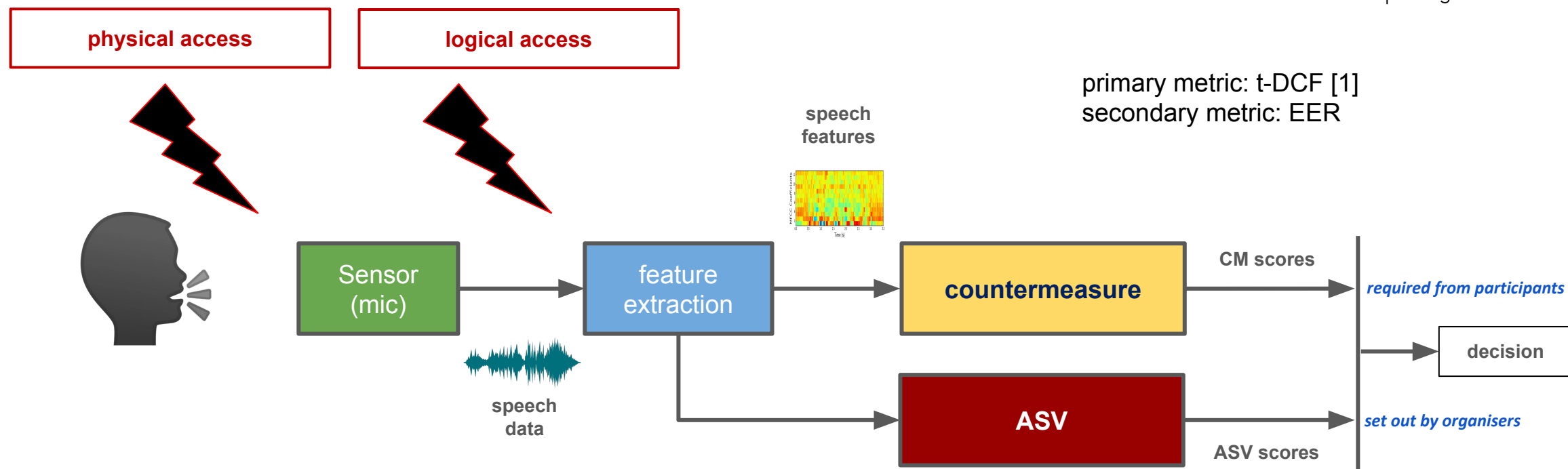
2021 - ~150 participants



ASVspoof 2019 database



- ASV centered
- logical access (LA) and physical access (PA) (separately evaluated)



1. Kinnunen T., Lee K., Delgado H., Evans N., Todisco M., Sahidullah Md., Yamagishi J., and Reynolds D. A., “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.

ASVspoof 2019 database



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

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

ASVspoof 2019 database



- ASVspoof 2019 [1] - logical and physical access
- common database and base protocol



Train

20 speakers

LA → 4 TTS and 2 VC attacks

PA → 27 acoustic / 9 replay configurations

Development

10 speakers

LA → 4 TTS and 2 VC attacks

PA → 27 acoustic / 9 replay configurations

Evaluation

unknown attacks

48 speakers

LA → 10 TTS and 3 VC attacks

PA → 27 acoustic / 9 replay configurations

1. Wang X. et al, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," Computer Speech & Language, 2020.

- attacks






	Input	Input processor	Duration	Conversion	Speaker represent.	Outputs	Waveform generator	Post process
A01	Text	NLP	HMM	AR RNN*	VAE*	MCC, F0	WaveNet*	
A02	Text	NLP	HMM	AR RNN*	VAE*	MCC, F0, BAP	WORLD	
A03	Text	NLP	FF*	FF*	One hot embed.	MCC, F0, BAP	WORLD	
A04	Text	NLP	-	CART	-	MFCC, F0	Waveform concat.	
A05	Speech (human)	WORLD	-	VAE*	One hot embed.	MCC, F0, AP	WORLD	
A06	Speech (human)	LPCC/MFCC	-	GMM-UBM	-	LPC	Spectral filtering + OLA	
A07	Text	NLP	RNN*	RNN*	One hot embed.	MCC, F0, BA	WORLD	GAN*
A08	Text	NLP	HMM	AR RNN*	One hot embed.	MCC, F0	Neural source-filter*	
A09	Text	NLP	RNN*	RNN*	One hot embed.	MCC, F0	Vocaine	
A10	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*	Mel-spectrograms	WaveRNN*	
A11	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*	Mel-spectrograms	Griffin-Lim [13]	
A12	Text	NLP	RNN*	RNN*	One hot embed.	F0+linguistic features	WaveNet*	
A13	Speech (TTS)	WORLD	DTW	Moment matching*	-	MCC	Waveform filtering	
A14	Speech (TTS)	ASR*	-	RNN*	-	MCC, F0, BAP	STRAIGHT	
A15	Speech (TTS)	ASR*	-	RNN*	-	MCC, F0	WaveNet*	
A16	Text	NLP	-	CART	-	MFCC, F0	Waveform concat.	
A17	Speech (human)	WORLD	-	VAE*	One hot embed.	MCC, F0	Waveform filtering	
A18	Speech (human)	MFCC/i-vector	-	Linear	PLDA	MFCC	MFCC vocoder	
A19	Speech (human)	LPCC/MFCC	-	GMM-UBM	-	LPC	Spectral filtering + OLA	

ASVspoof 2019 LA



- speech samples



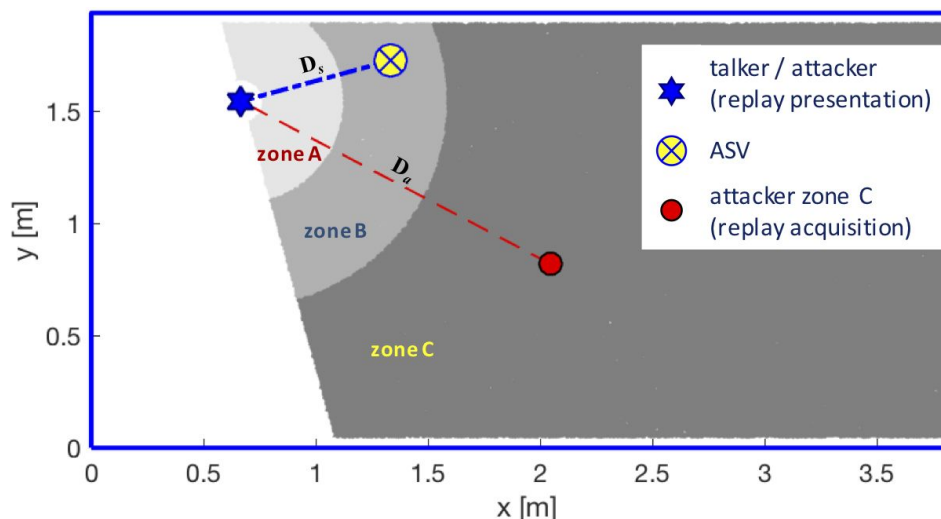
Bona fide	Attack A10 (Google Tacotron 2)
	
	
	



- 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 .

- acoustic environment
 - room size
 - convolutive noise
 - recording distance between bonafide user and ASV
- replay acquisition
 - recording distance between bonafide user and attacker
- replay presentation
 - device quality (loudspeaker)



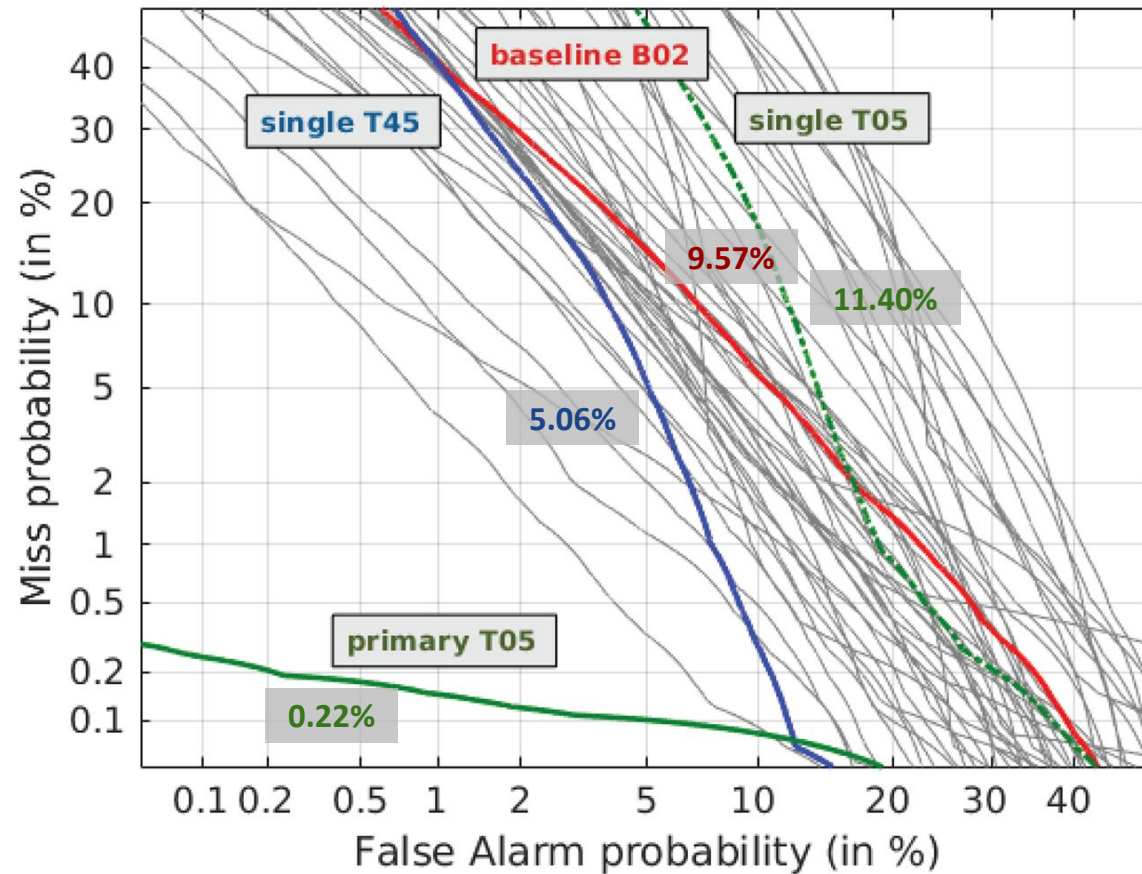
Environment definition	labels		
	a	b	c
S: Room size (m ²)	2-5	5-10	10-20
R: T60 (ms)	50-200	200-600	600-1000
D _s : Talker-to-ASV distance (cm)	10-50	50-100	100-150

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

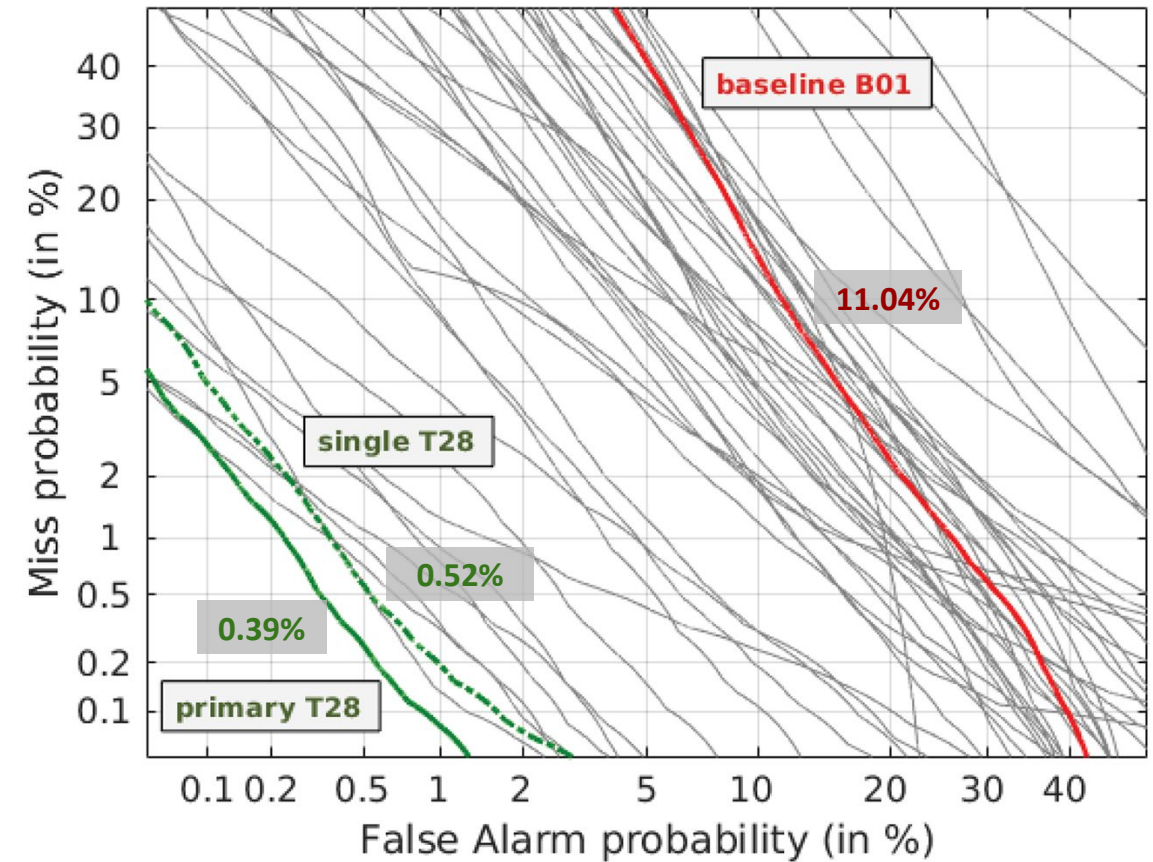
ASVspoof 2019 challenge results

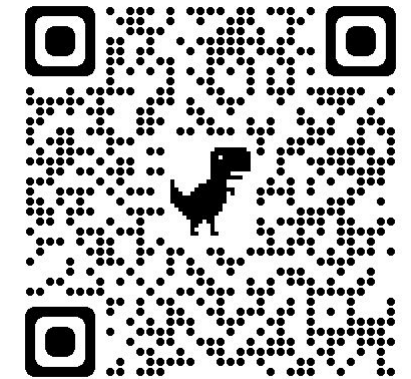


LA



PA





- **logical access** (LA): bona fide and spoofed utterances generated using text-to-speech (TTS) and voice conversion (VC) algorithms are communicated across telephony and VoIP networks with various coding and transmission effects
- **physical access** (PA): bona fide utterances are made in a real, physical space in which spoofing attacks are captured and then replayed within the same physical space using replay devices of varying quality
- **speech deepfake** (DF): a fake audio detection task comprising bona fide and spoofed utterances generated using TTS and VC algorithms. Similar to the LA task (includes compressed data) but without speaker verification.

ASVspoof 2021 Workshop



- The **ASVspoof 2021 Workshop**, an official Interspeech 2021 satellite event, will be held online in the form of a Zoom Webinar on **September 16th, 2021**.
- Participation is free of charge and open to all, but registration is mandatory →
- ... or visit <https://www.asvspoof.org/workshop>

