The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection

Tomi Kinnunen, University of Eastern Finland, FINLAND

Md Sahidullah, University of Eastern Finland, FINLAND

Héctor Delgado, EURECOM, FRANCE

Massimiliano Todisco, EURECOM, FRANCE

Nicholas Evans, EURECOM, FRANCE

Junichi Yamagishi, Univ. of Edinburgh, UK & National Institute of Informatics, JAPAN

Kong Aik Lee, Institute for Infocomm Research, SINGAPORE

Organizers

Tomi H. Kinnunen

UEF, Finland

Md Sahidullah

UEF, Finland

Héctor Delgado

EURECOM, France

Massimiliano Todisco

EURECOM, France

Nicholas Evans

EURECOM, France

Junichi Yamagishi

Univ. of Edinburgh, UK

NII, Japan

Kong Aik Lee

I^2R, Singapore

Structure of the session

First slot 11:00 – 13:00 CHAIRS: Tomi Kinnunen, Junichi Yamagishi

INTRODUCTION, 30 mins

6 ORAL PRESENTATIONS, each 12 + 3 min

Second slot 14:30 – 16:30 CHAIRS: Nicholas Evans, Kong Aik Lee

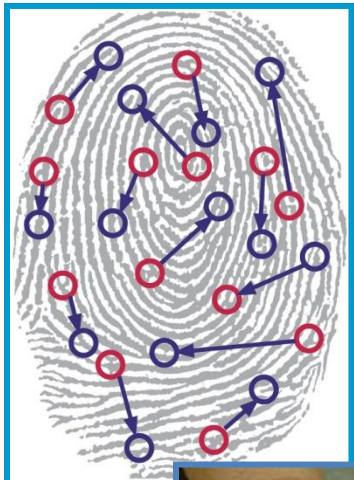
6 ORAL PRESENTATIONS, each 12 + 3 min

GENERAL DISCUSSION @ 16:00---

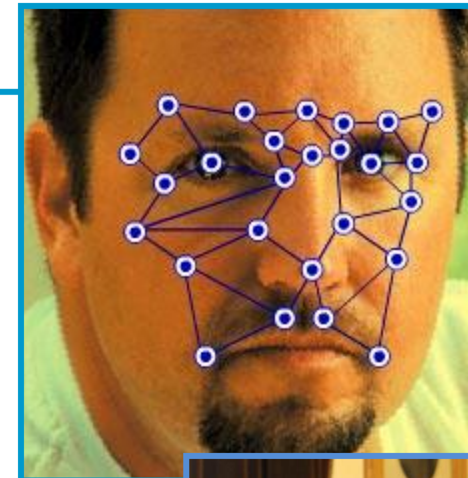
Spoofing attacks

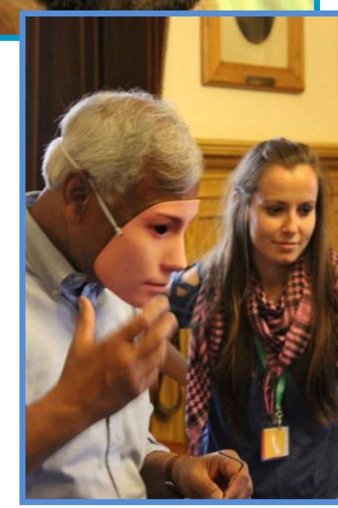
a.k.a. presentation attacks [ISO/IEC 30107-1:2016]

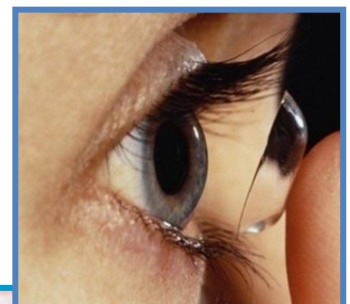
Finger-print

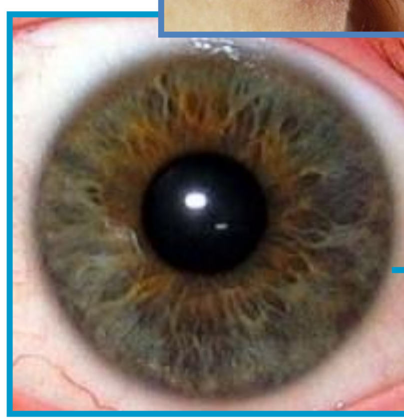


Face



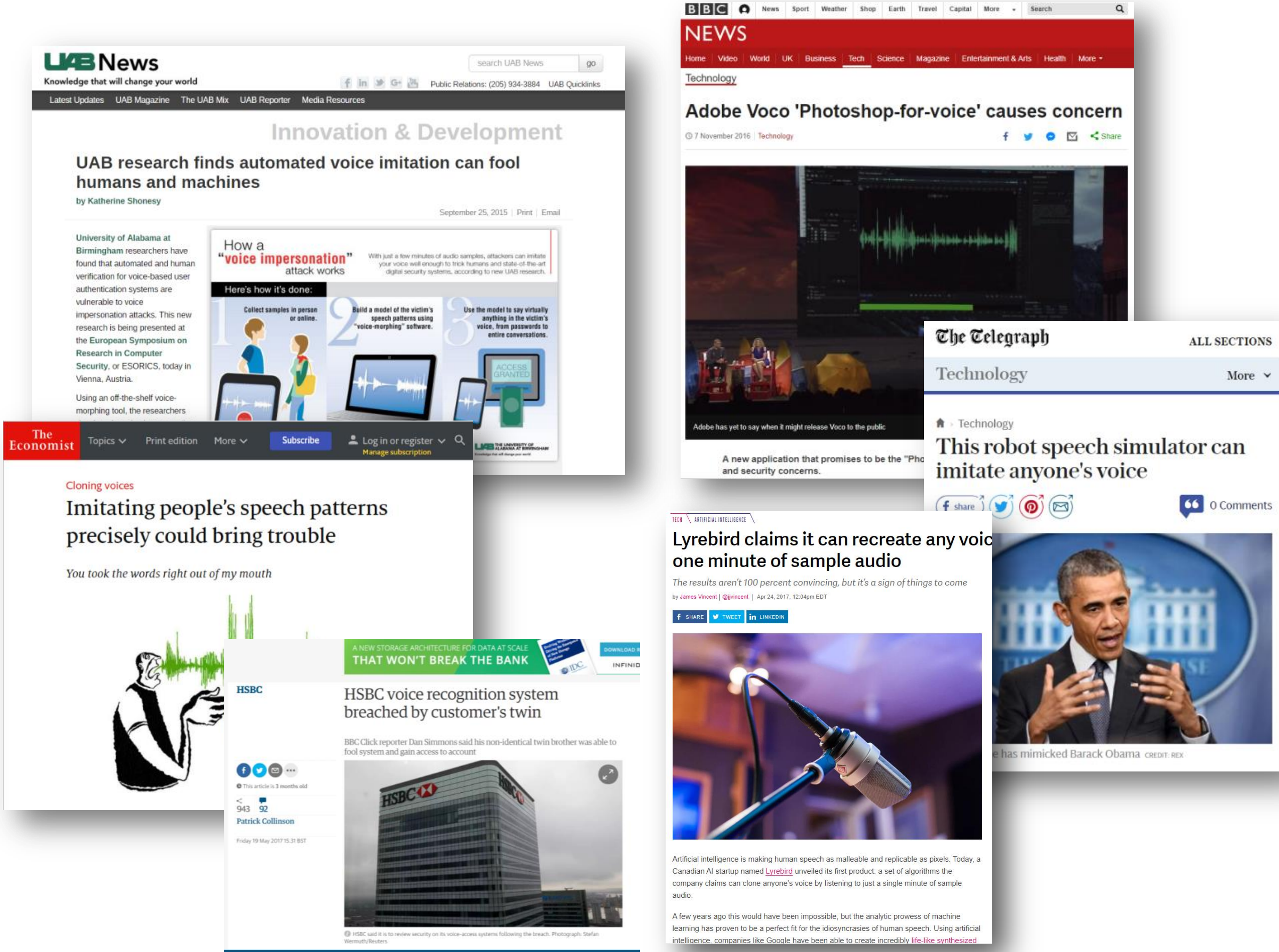






Iris

Sources: unknown



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 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

How a "voice impersonation" attack works

With just a few minutes of audio samples, attackers can imitate your voice well enough to trick humans and state-of-the-art digital security systems, according to new UAB research.

Here's how it's done:

1 Collect samples in person or online.

2 Build a model of the victim's speech patterns using "voice-morphing" software.

3 Use the model to say virtually anything in the victim's voice, from passwords to entire conversations.

ACCESS GRANTED

BBC 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

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

A new application that promises to be the "Pho... and security concerns.

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

The Telegraph | ALL SECTIONS

Technology | More

Technology

This robot speech simulator can imitate anyone's voice

share | 0 Comments

...e has mimicked Barack Obama CREDIT: REX

HSBC

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

This article is 3 months old

943 | 92

Patrick Collinson

Friday 19 May 2017 15.31 BST

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

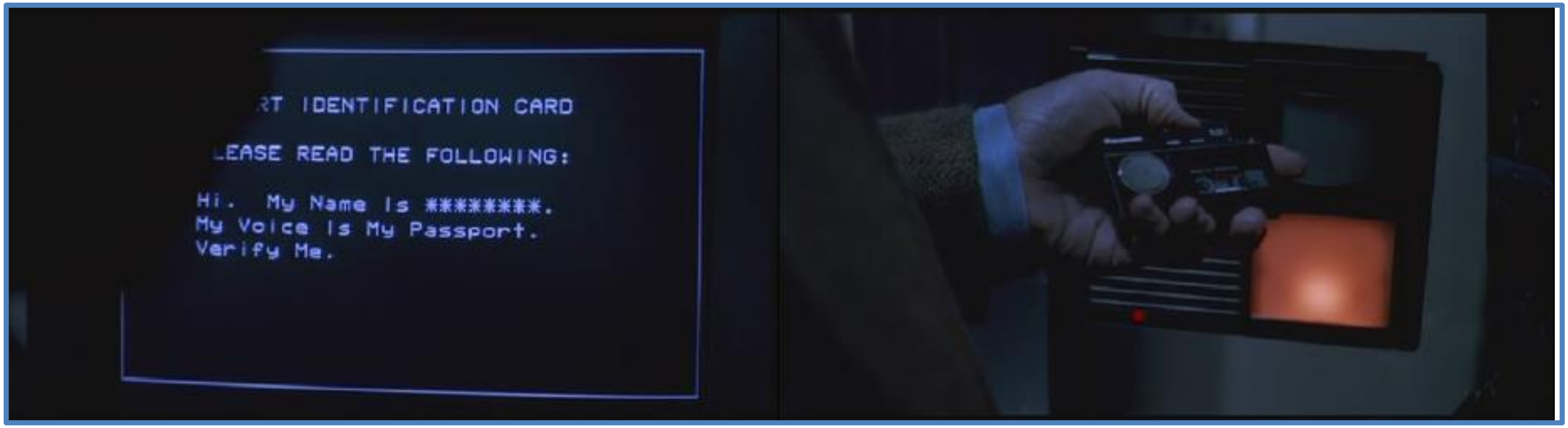
SHARE | TWEET | LINKEDIN

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.

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

Replay attack

replay spoofing – Sneakers (1992)



Universal Pictures

History of ASVspoof

OCTAVE
Objective Control of Talker Verification

1999 | 2006 | 2014 | 2016 | 2017

OCTAVE project starts

small, purpose collected datasets

2013 Interspeech special session

adapted, standard datasets

INTERSPEECH 2017
Situated Interaction
August 20-24, 2017 | Stockholm, Sweden
www.interspeech2017.org

common datasets, metrics, protocols

common datasets, replay, generalisation, channel variation

ASVspoof 2015

ASVspoof 2017

Replay attack countermeasures

1. Phrase prompting with utterance verification

Did the user speak the prompted text ?

Can be circumvented using voice conversion

2. Audio fingerprinting

Do I know this recording ?

Dynamically increasing database size

3. Speaker-independent replay detection

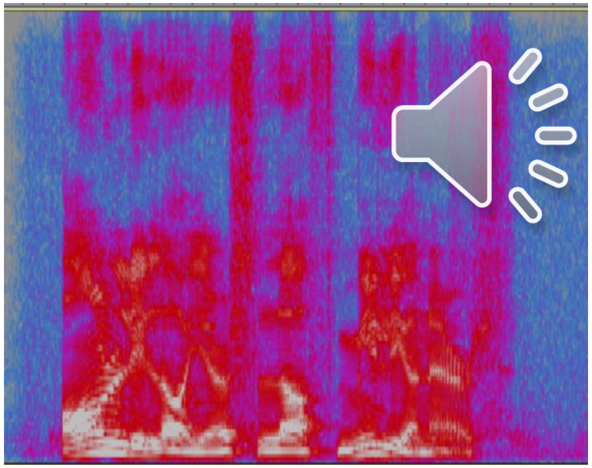
Is this recording authentic or replayed one ?

Most general - but can it be done?

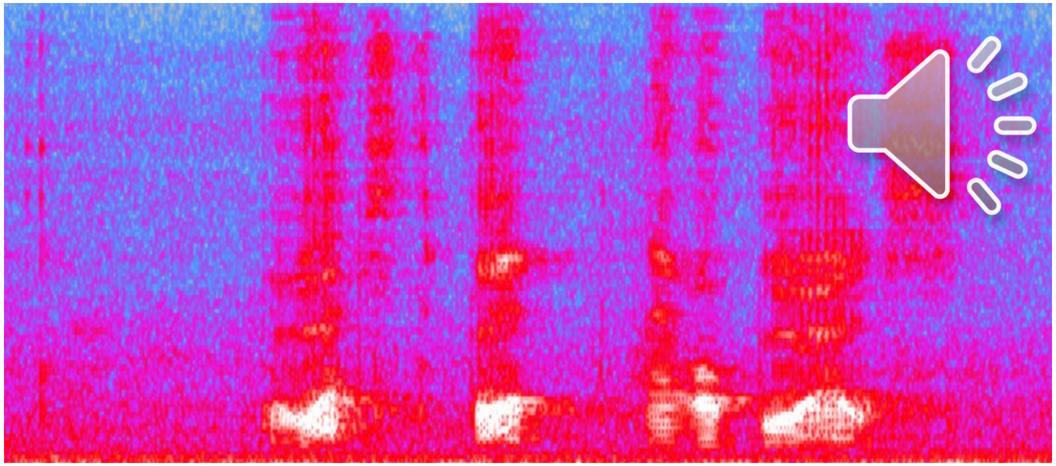
ASVspoof 2017

1. T. Stafylakis, M. J. Alam, and P. Kenny, "Text dependent speaker recognition with random digit strings," IEEE/ACM T-ASLP 24(7): 1194–1203, 2016.
2. Q. Li, B.-H. Juang, and C.-H. Lee, "Automatic verbal information verification for user authentication," IEEE Transactions on Speech and Audio Processing, vol. 8, no. 5, pp. 585–596, Sep 2000.
3. T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. Sarkar, N. B. Thomsen, V. Hautamaki, N. Evans, and Z.-H. ¨ Tan, "Utterance verification for text-dependent speaker recognition: a comparative assessment using the RedDots corpus," Proc. INTERSPEECH, 2016
4. C. Ouali, P. Dumouchel, and V. Gupta, "A robust audio fingerprinting method for content-based copy detection," in Proc. 12th International Workshop on Content-Based Multimedia Indexing (CBMI), June 2014, pp. 1–6
5. M. Malekesmaeili and R. Ward, "A local fingerprinting approach for audio copy detection," Signal Processing, vol. 98, pp. 308 – 321, 2014

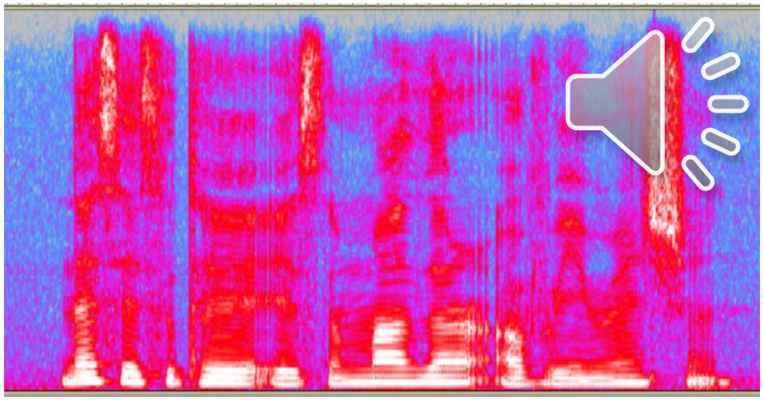
Replayed or nonreplayed ?



Authentic (non-replayed)



Replayed

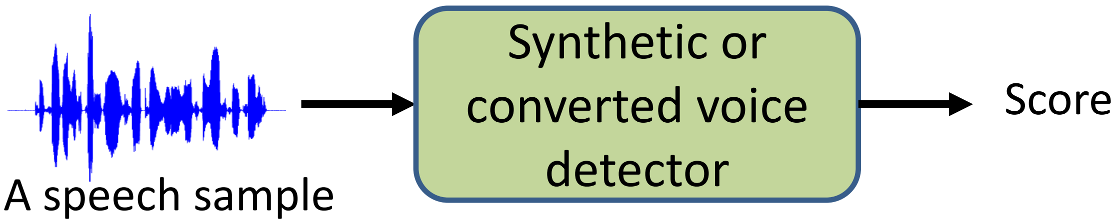


Replayed

ASVspoof challenge task

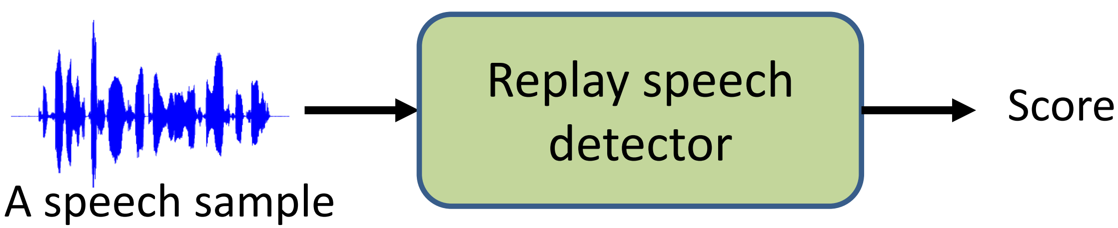
Standalone, speaker-independent detection of spoofing attacks

ASVspoof 2015



High score → more likely a live human being
Low score → more likely a spoofed sample

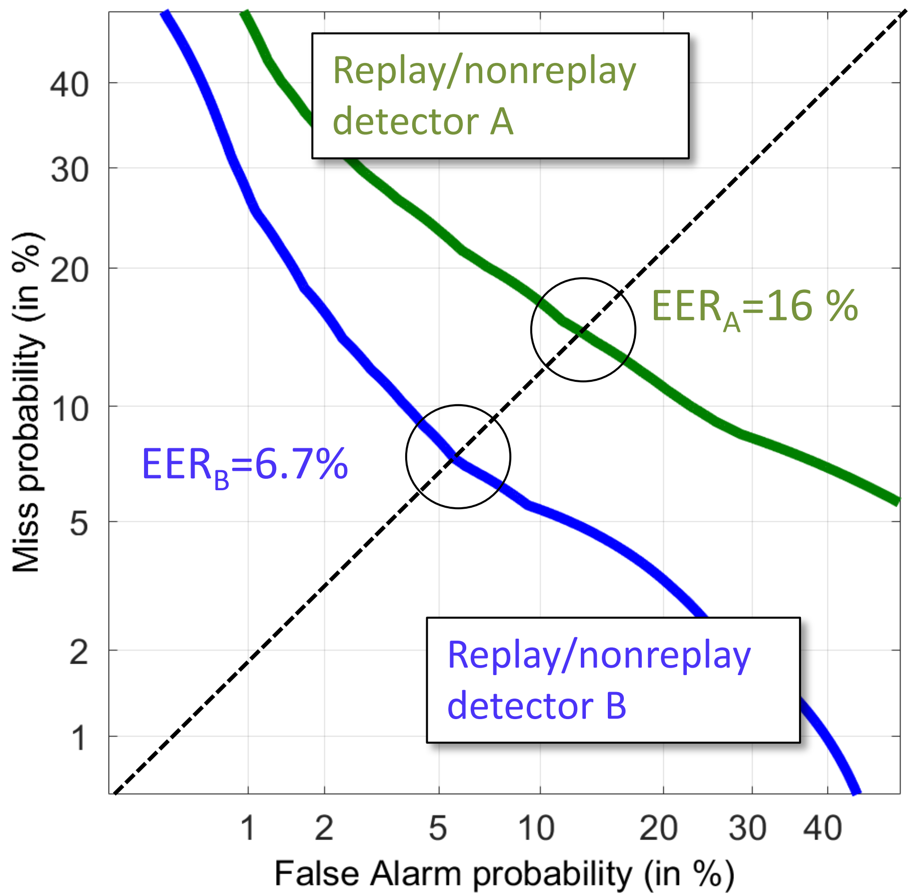
ASVspoof 2017



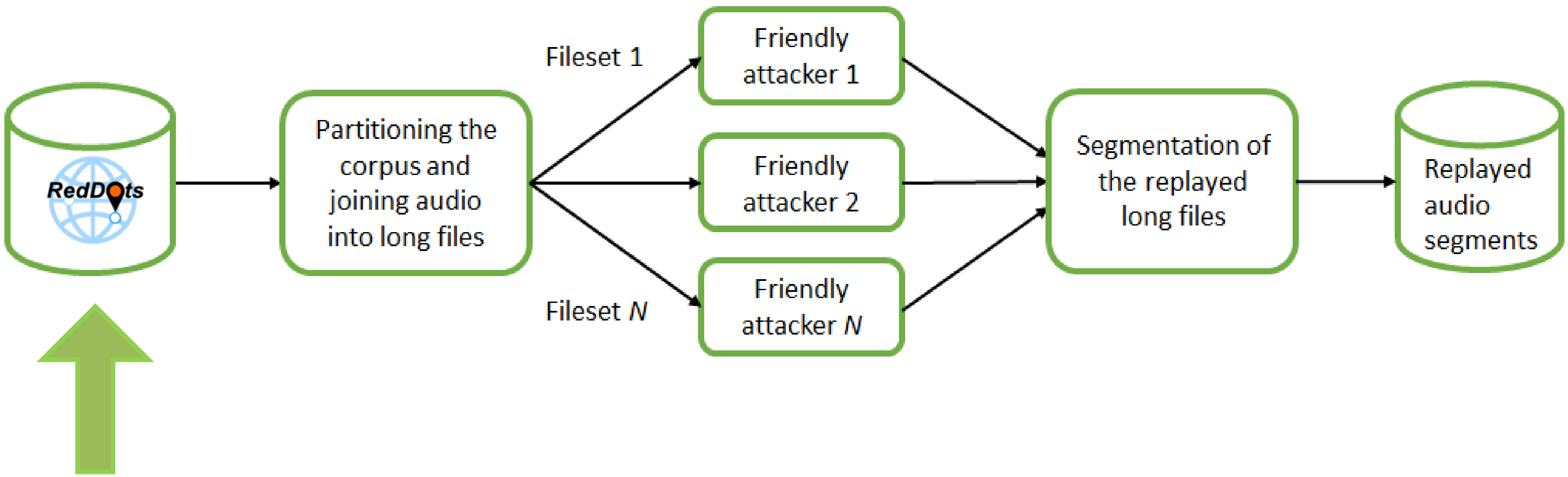
Evaluation metric:

Equal error rate (EER) of replay-nonreplay discrimination

- **ASVspoof 2015**: EERs averaged across attacks
- **ASVspoof 2017**: EERs from pooled scores



Crowdsourced replay attacks



RedDots corpus

[https://sites.google.com/site/thereddotsproject/]

- Text-dependent automatic speaker verification
- Collected by volunteers (ASV researchers)
- Various Android devices, speakers, accents

Examples of replay configurations

REPLAY CONFIGURATION =
Playback device + Environment +
Recording device

Smartphone → Smartphone

Headphones
→ PC mic

High-quality loudspeaker
→ smartphone, anechoic room

High-quality loudpspeaker
→ high-quality mic

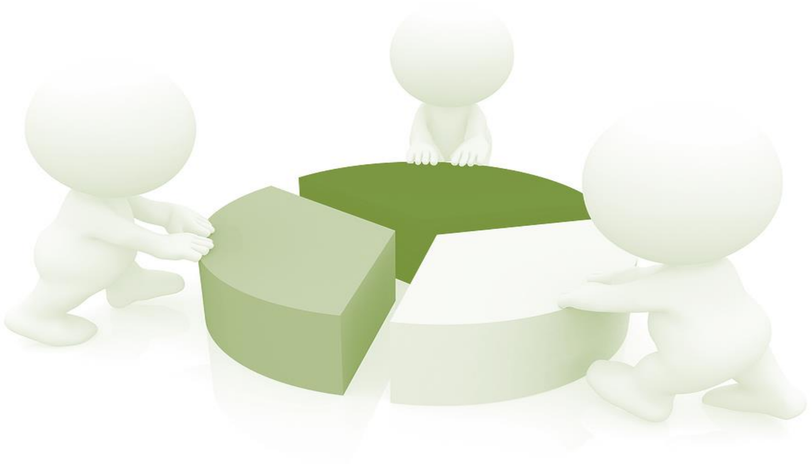
Laptop line-out
→ PC line-in using a cable

T. Kinnunen et al., "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5395-5399.

TRAINING SET

- 10 speakers
- 3 replay configs

- Ground truth provided
- Re-partitioning allowed



DEVELOPMENT SET

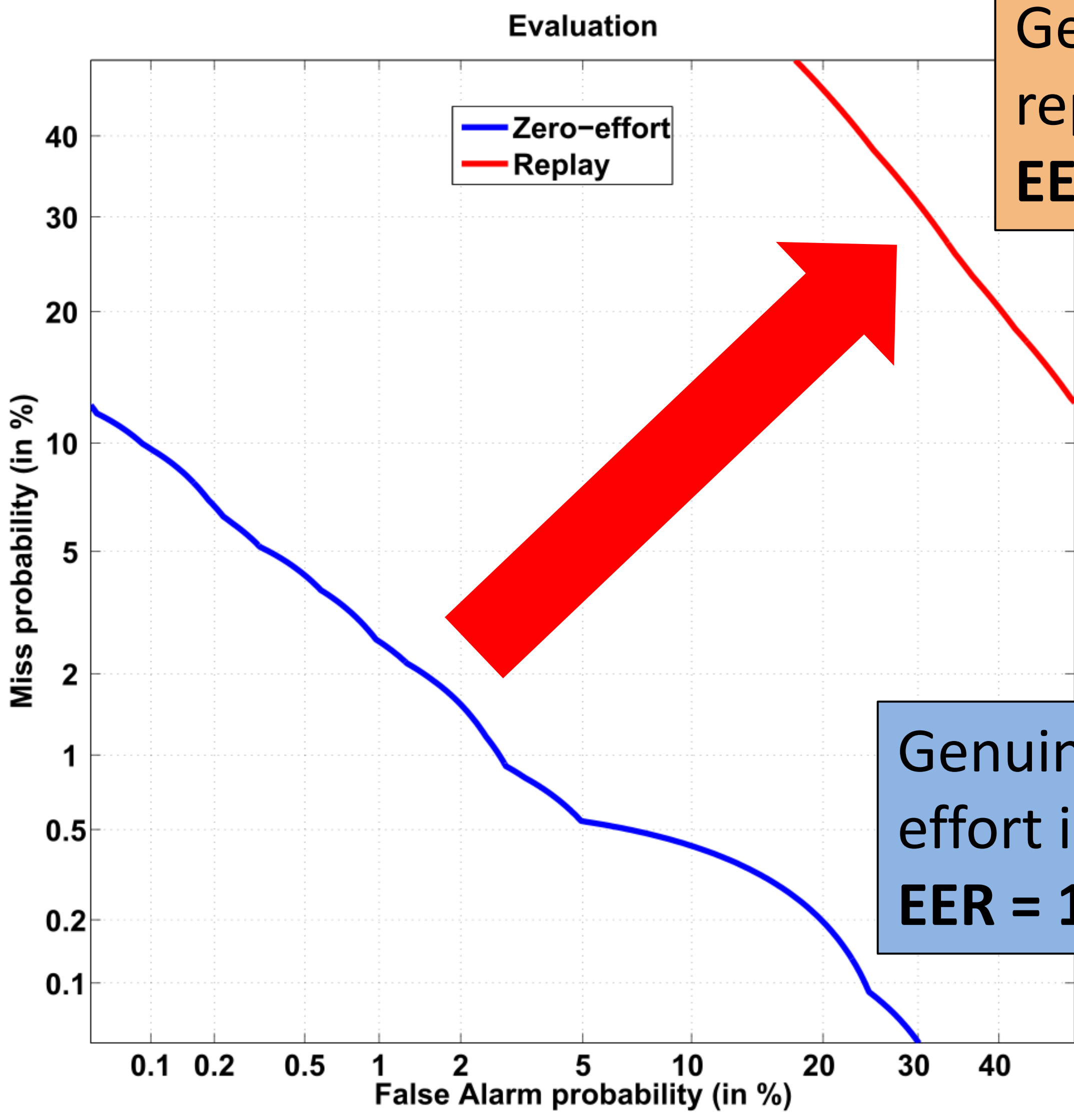
- 8 speakers
- 10 replay configs

EVAL SET

- 24 speakers
- 110 replay configs

Impact of replay samples to ASV

gmm-ubm system



Genuine vs. replay impostors
EER = 31.5 %

Genuine vs. zero-effort impostors
EER = 1.8 %

Participant statistics

- Registration: 113 teams or individuals
- Submitted results: 49 (43%)

Challenge results and further analyses

- Official challenge results

- Further analyses

Official challenge results

Common primary submissions' results

- **Very difficult challenge!**
- 21 submissions outperformed the baseline
- S01: **>70% relative improvement w.r.t baseline B01**
- B01 – B02: Important performance improvement when using **pooled train+dev** data for training

Sxx: Regular submission
Bxx: Baseline system
Dxx: Late submission

Summary of top 10 systems

ID	EER	Features	Post-proc.	Classifiers	Fusion	#Subs.	Training
S01	6.73	Log-power Spectrum, LPCC	MVN	**CNN**, **GMM**, **TV**, **RNN**	Score	3	T
S02	12.34	**CQCC**, MFCC, PLP	WMVN	**GMM-UBM**, **TV-PLDA**, **GSV-SVM**, **GSV-GBDT**, **GSV-RF**	Score	-	T
S03	14.03	MFCC, IMFCC, RFCC, LFCC, PLP, **CQCC**, SCMC, SSFC	-	**GMM**, **FF-ANN**	Score	18	T+D
S04	14.66	RFCC, MFCC, IMFCC, LFCC, SSFC, SCMC	-	**GMM**	Score	12	T+D
S05	15.97	Linear filterbank feature	MN	**GMM**, **CT-DNN**	Score	2	T
S06	17.62	**CQCC**, IMFCC, SCMC, Phrase one-hot encoding	MN	**GMM**	Score	4	T+D
S07	18.14	HPCC, **CQCC**	MVN	**GMM**, **CNN**, **SVM**	Score	2	T+D
S08	18.32	IFCC, CFCCIF, Prosody	-	**GMM**	Score	3	T
S10	20.32	**CQCC**	-	**ResNet**	None	1	T
S09	20.57	SFFCC	-	**GMM**	None	1	T
D01	7.00	MFCC, **CQCC**, WT	MVN	**GMM**, **TV-SVM**	Score	26	T+D

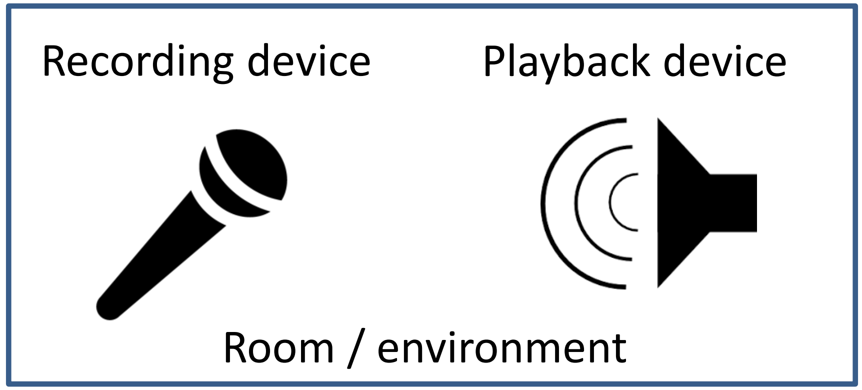
Using baseline CQCC features

DNN-based classifier
Other classifier

T: training
T+D: training + development

Further analyses

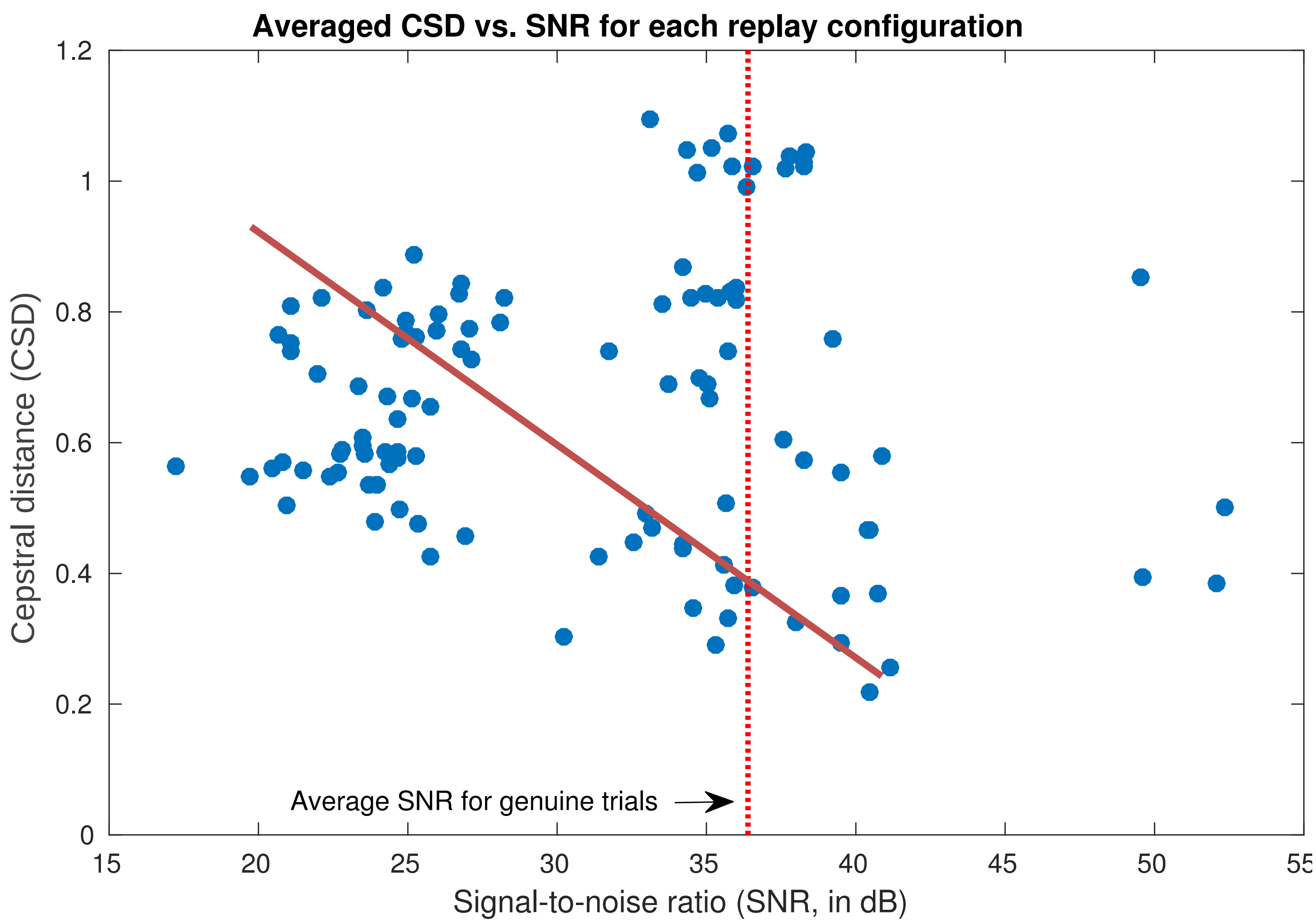
Defining evaluation conditions



REPLAY CONFIGURATION

- 110 replay configurations in evaluation set
- Characterize replay configurations through objective measurements
  - **Signal-to-noise ratio (SNR)**
  - **Cepstral distance (CSD)**: measures the degradation of a replayed recording w.r.t. its source recording
- Intuition:
  - More difficult attacks → **High SNR, low CSD**
  - Easier attacks → **Low SNR, high CSD**

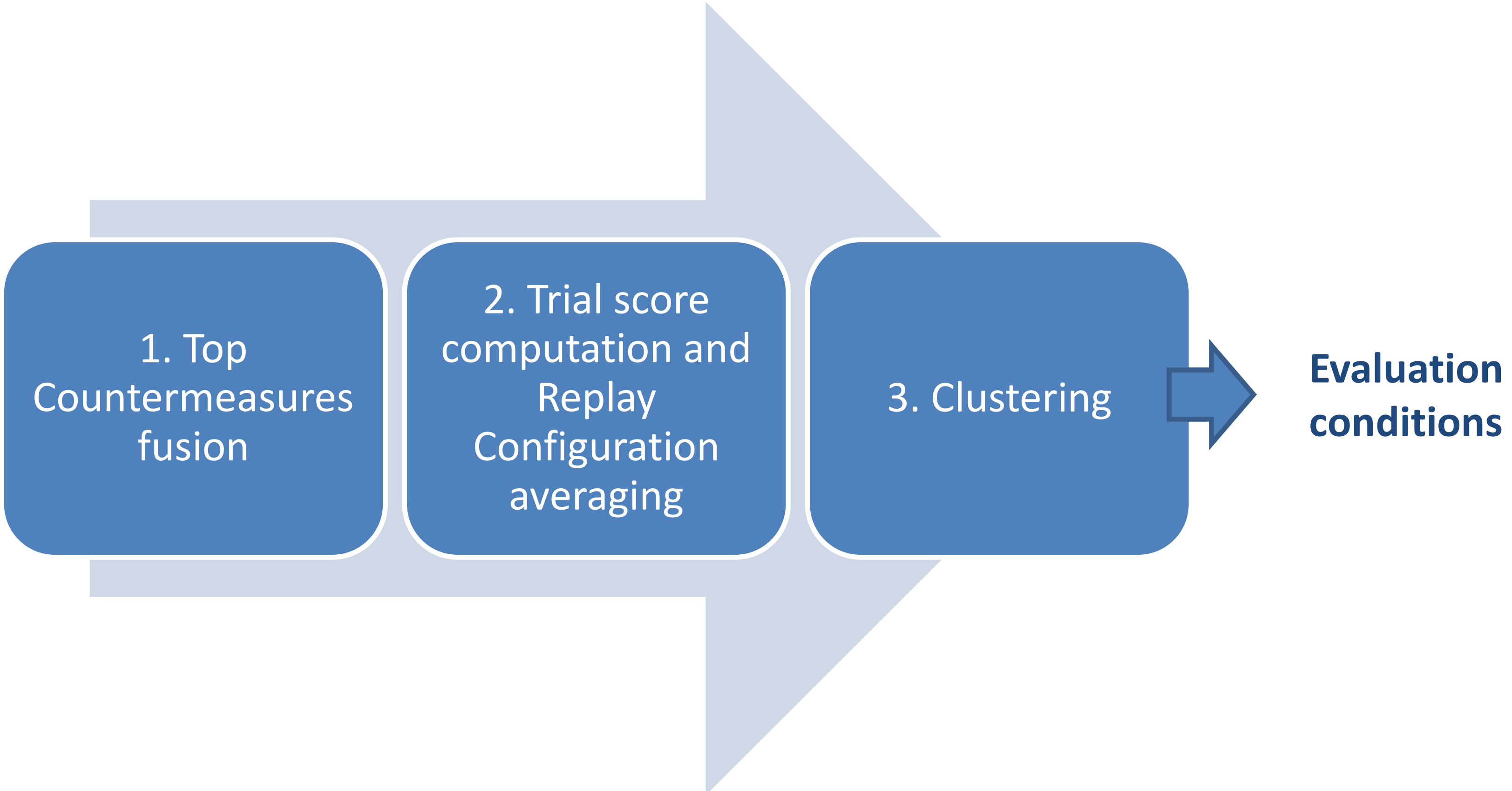
Average quality measures per replay configuration



Average CSD vs. SNR scatter plot for the 110 replay configurations

Data-driven clustering process

Alternative approach: define evaluation conditions according to countermeasure performance



Data-driven clustering process

1. Countermeasure fusion

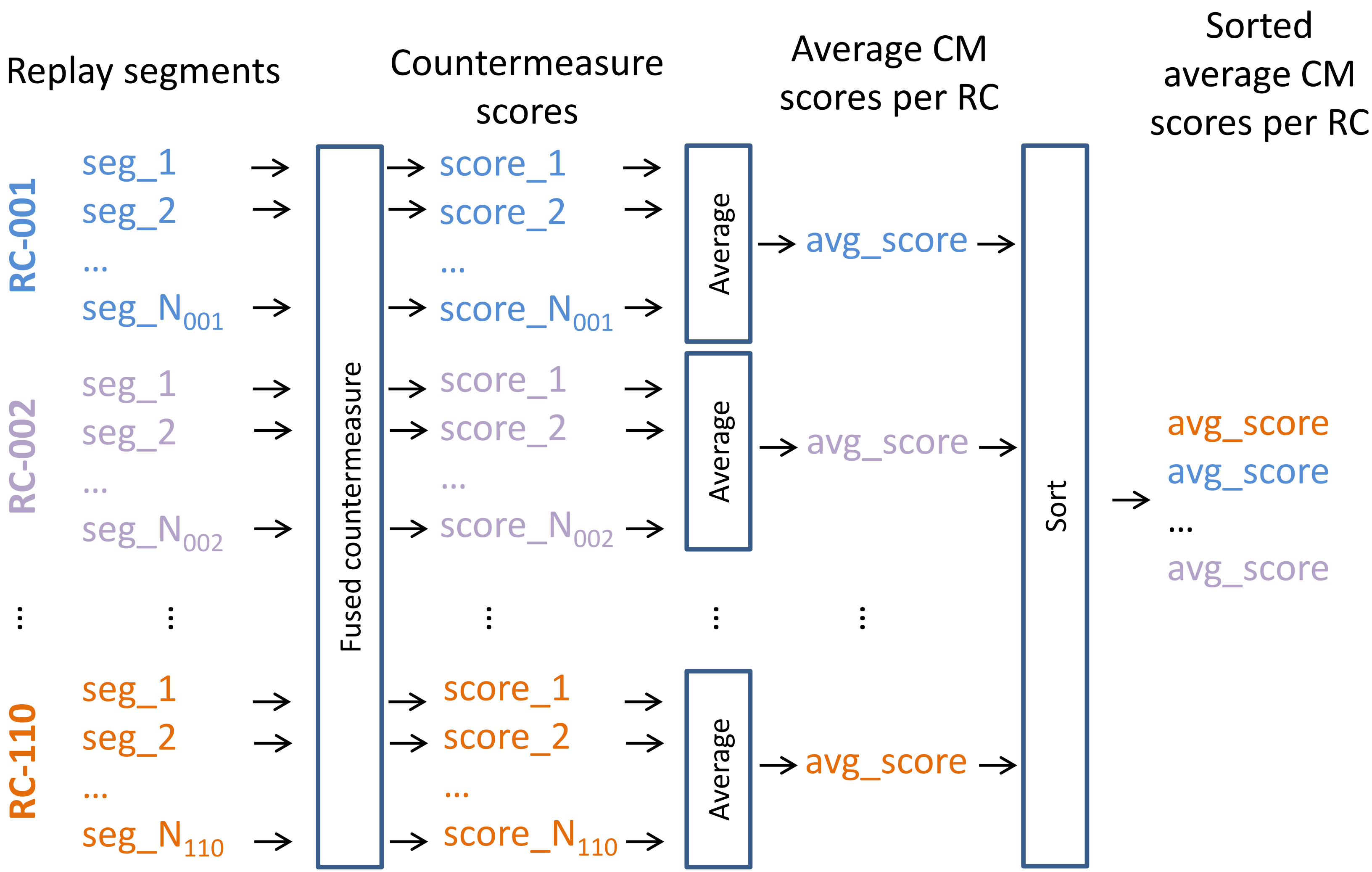
Oracle linear fusion[1] of systems S01 to B01 to obtain a high performance countermeasure

System	EER (%)
S01	6.73
S02	12.34
S03	14.03
S04	14.66
S05	15.97
S06	17.62
S07	18.14
S08	18.32
S10	20.32
S09	20.57
S11	21.11
S12	21.51
S13	21.98
S14	22.17
S15	22.39
S19	23.16
S18	23.24
S17	23.29
S10	23.78
B01	24.77
D01	7.00
Fused	**2.76**

[1]Using the Bosaris toolkit

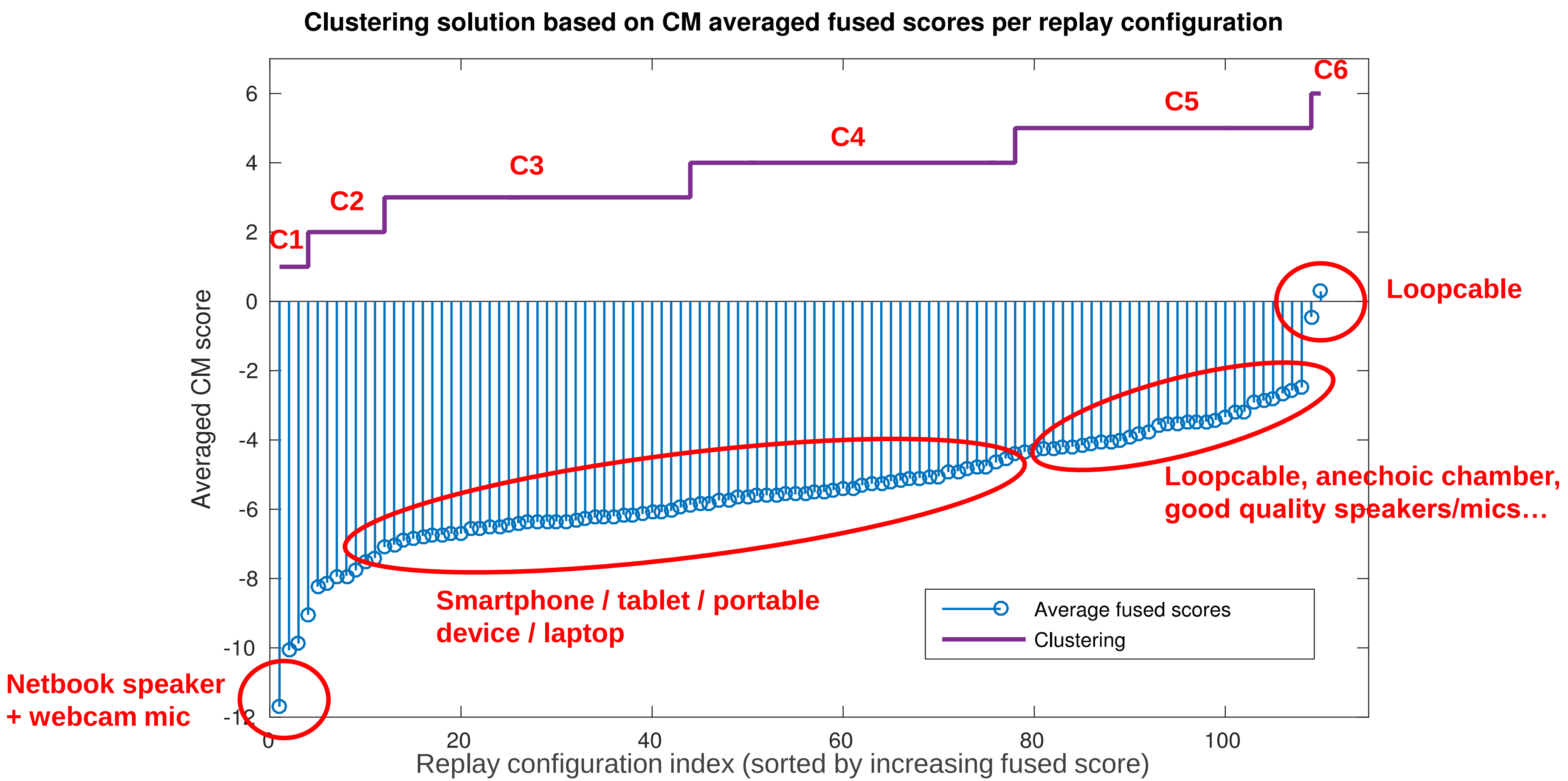
Data-driven clustering process

2. Average Replay Configuration (RC) scores computation and sorting

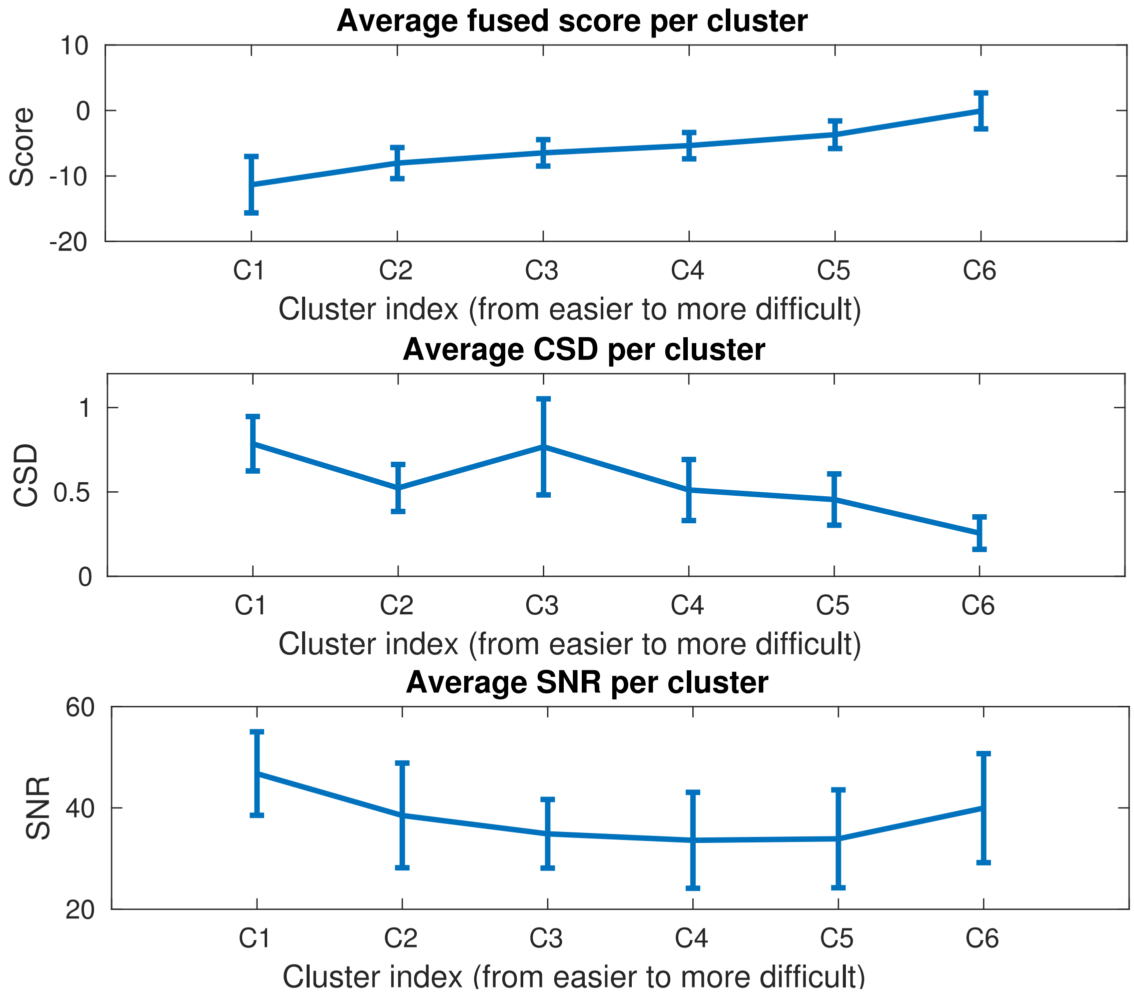


Data-driven clustering process

3. Average scores clustering with k-means

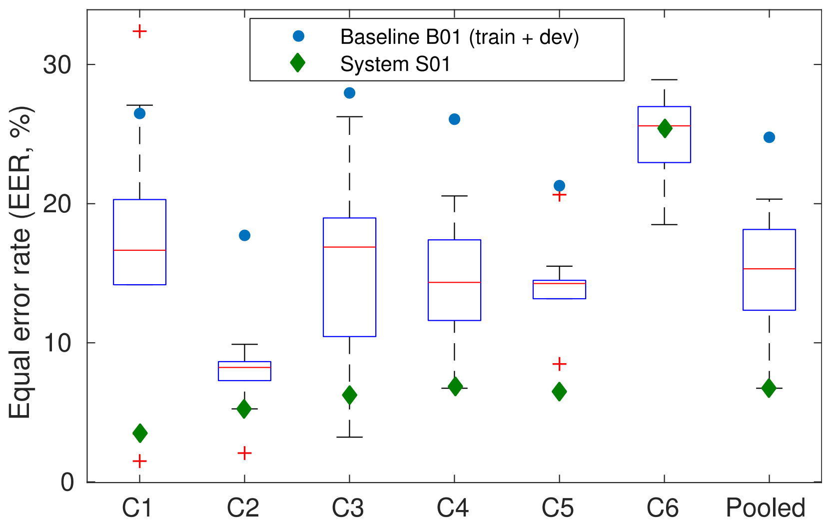


Obtained evaluation conditions

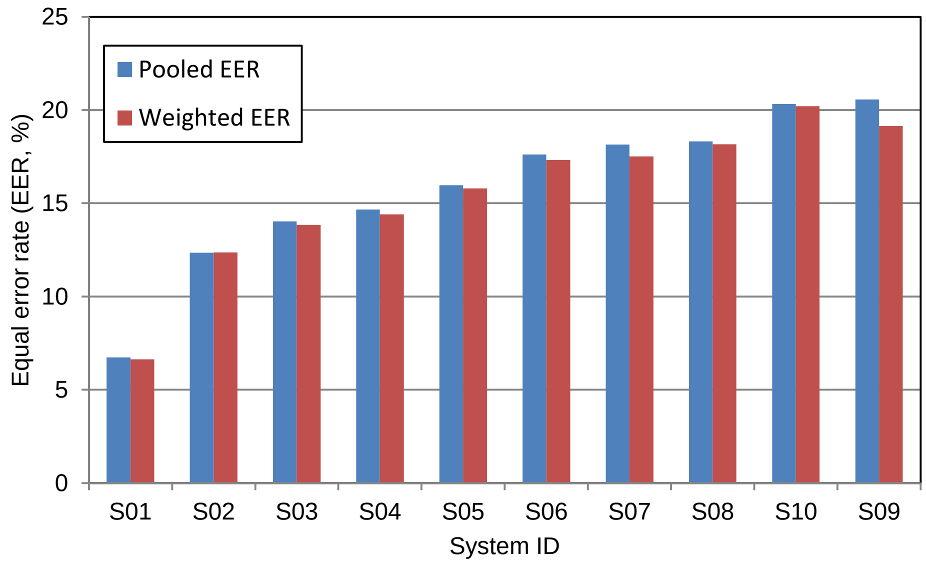


Averaged fused score, cepstral distortion and signal-to-noise ratio of the resulting evaluation conditions

Performance of top-10 primary submissions per evaluation condition



Box plot of top-10 systems' performance for clusters C1-C6

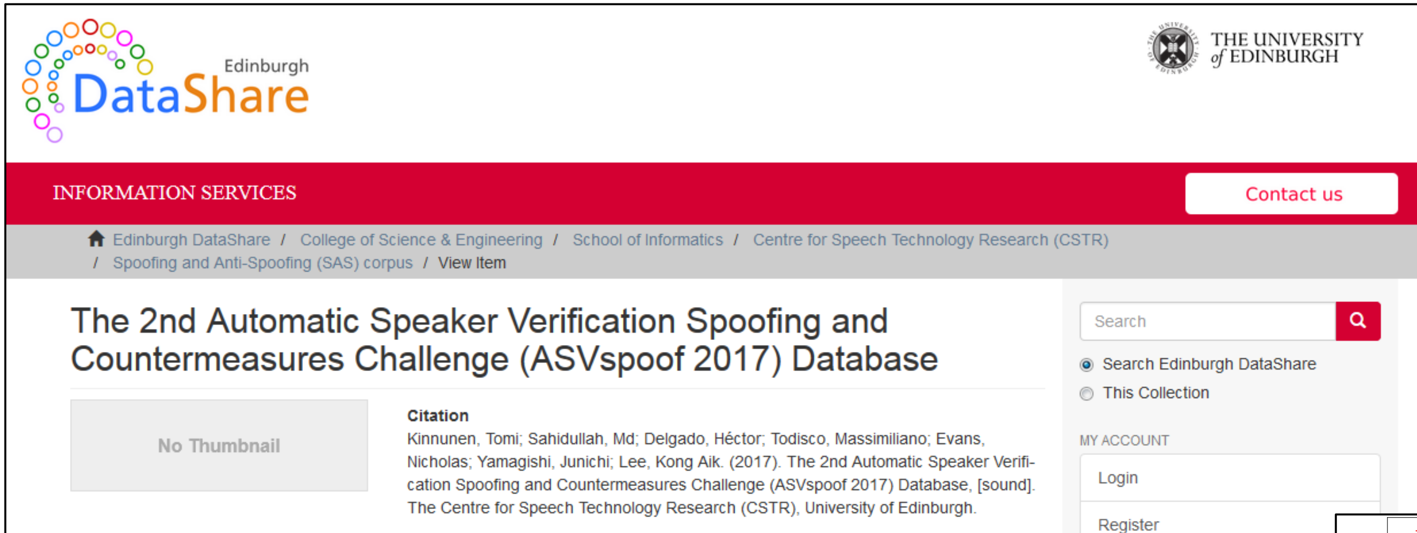


Pooled EER vs. weighted EER for top-10 systems

(equivalent to average EER used in ASVspoof 2015)

Conclusions

- Successful crowdsourcing approach to replay data collection
- Probably the most ‘wild’ replay data for ASV
  - Difficult to characterize
- Top-ranked system
  - ~70% relative improvement w.r.t. the baseline system
  - Fusion of only 3 subsystems!
- Encouraging performance
  - Limits of replay detection
  - Excepting unrealistic attacks (loopcable), high detection performance for high quality attacks



http://dx.doi.org/10.7488/ds/2105