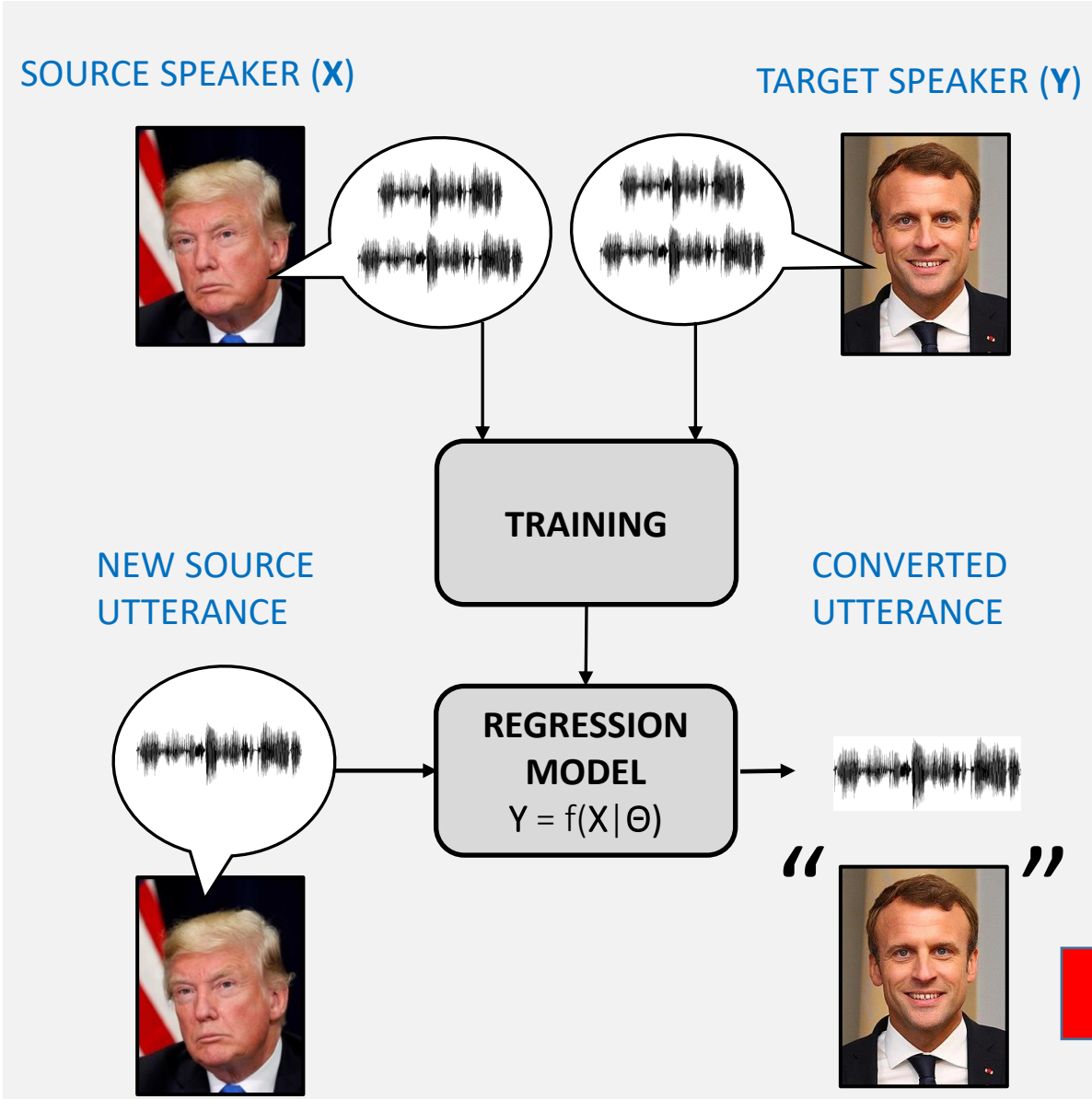
A Spoofing Benchmark for the 2018 Voice Conversion Challenge:
Leveraging from Spoofing Countermeasures for Speech Artifact Assessment

Tomi Kinnunen, Jaime Lorenzo-Trueba,
Junichi Yamagishi, Tomoki Toda, Daisuke
Saito, Fernando Villavicencio, Zhenhua Ling

http://www.vc-challenge.org/

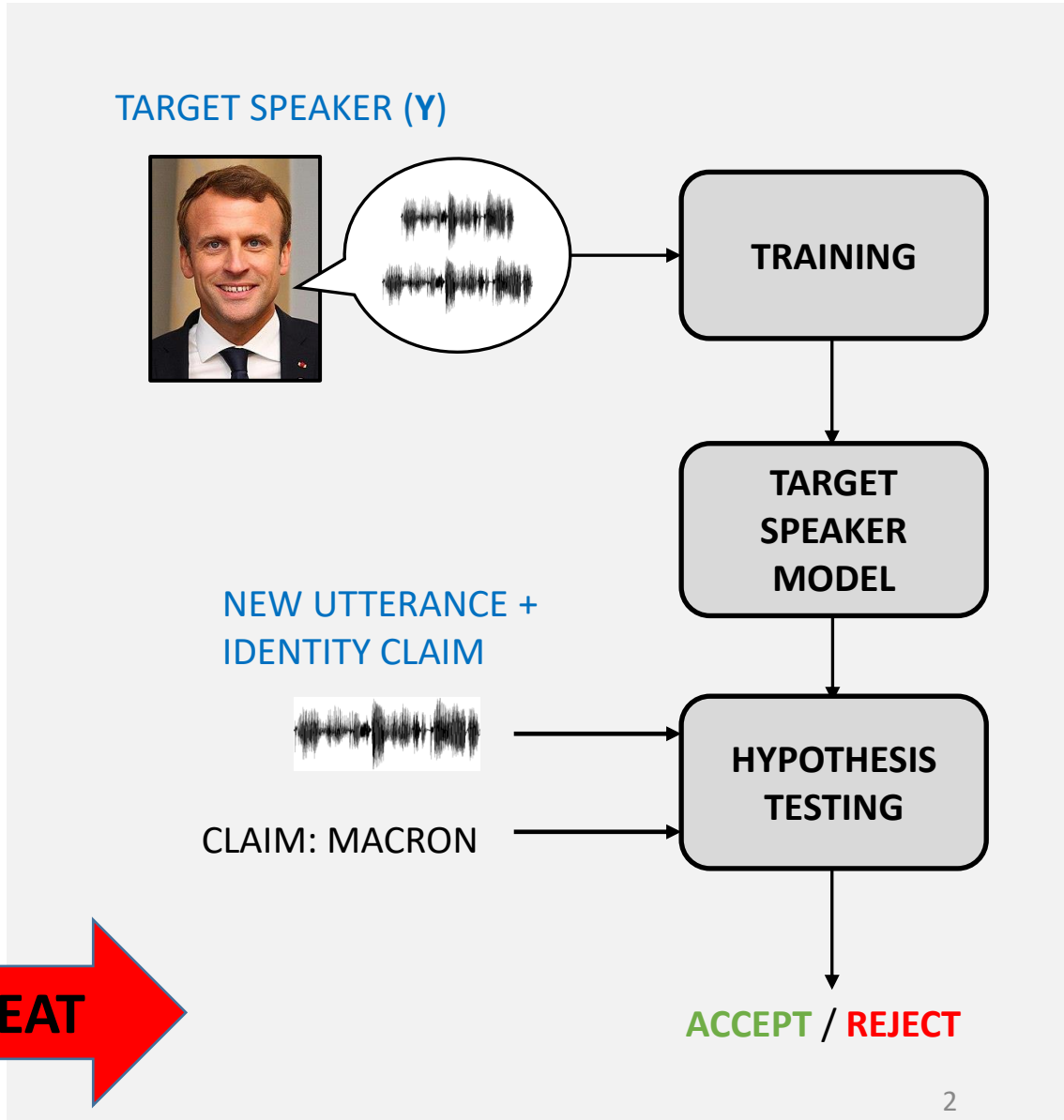
Voice conversion (VC)

Voice cloning by a computer



Automatic speaker verification (ASV)

Verifying person’s identity based on voice

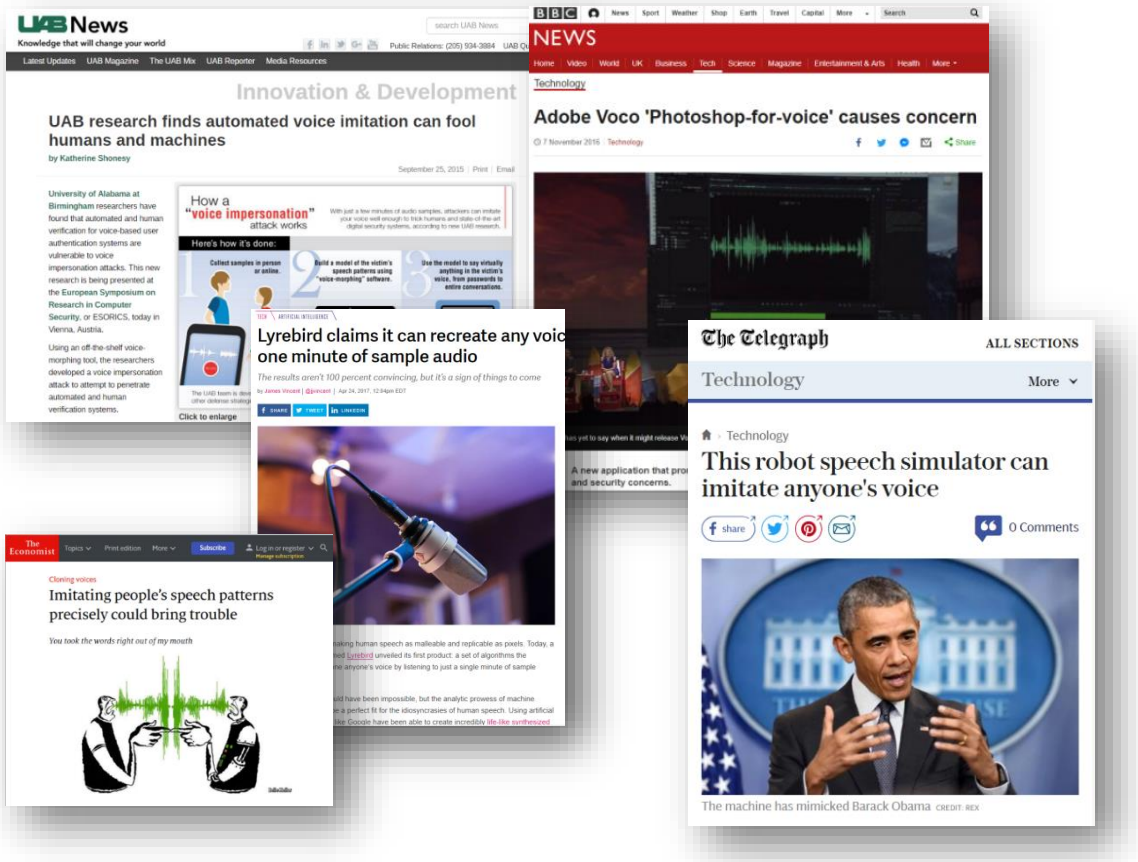


THREAT

Biometric spoofing attacks (presentation attacks)

- ISO standard: ISO/IEC 30107-1:2016
- Ratha, N.K., Connell, J.H., Bolle, R.M., 2001. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal* 40, 614– 634.
- Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, "Spoofing and Countermeasures for Speaker Verification: a Survey", *Speech Comm*, 66: 130--153, 2015

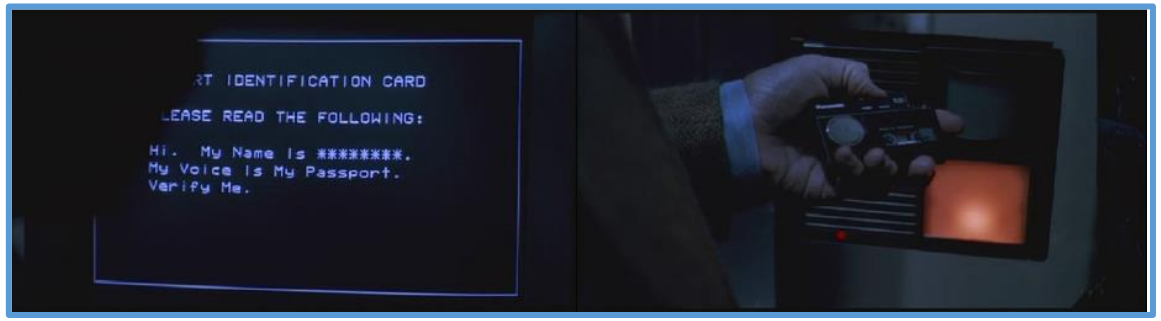
TEXT-TO-SPEECH AND VOICE CONVERSION

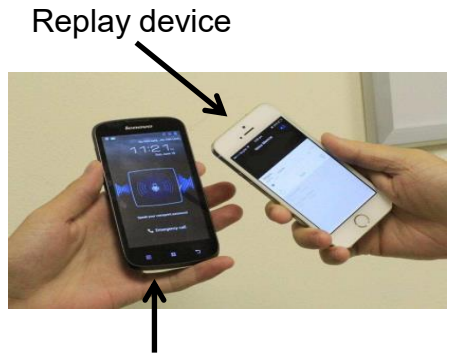


REPLAY

Sneakers (1992)

Universal Pictures

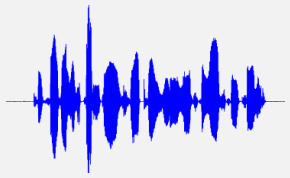




Spoofing countermeasure

System to detect presence of ”fake” speech

$H_0 : \mathcal{Y}$ presents human speech
$H_1 : \mathcal{Y}$ is machine-generated (artificial) or modified

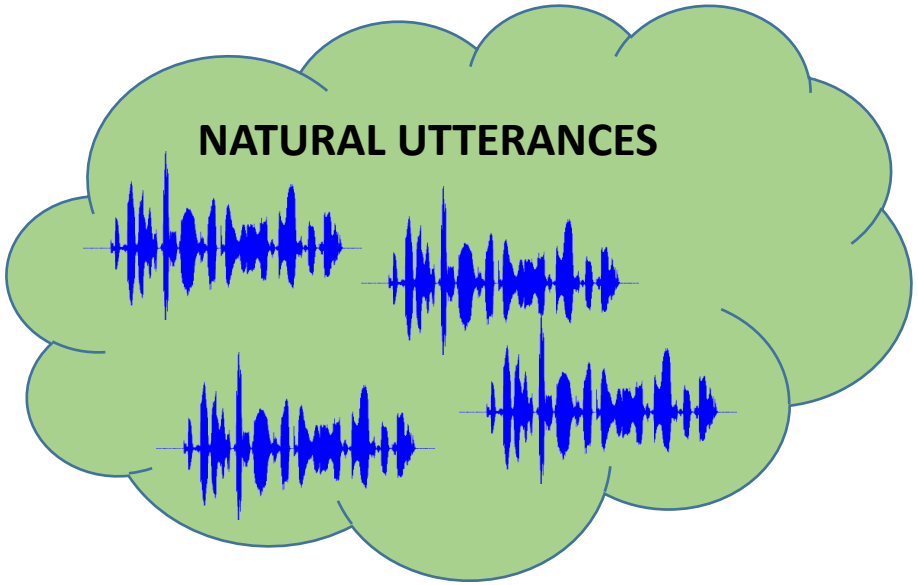


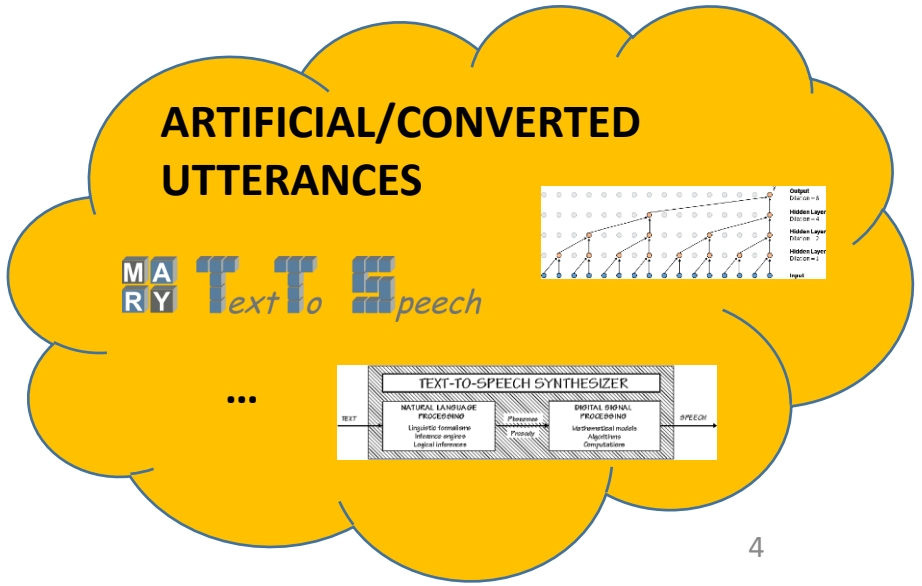
SPOOFING COUNTERMEASURE

HUMAN
or
SPOOF

Utterance
$\mathcal{Y}$

$$\ell = \log \frac{p(\mathcal{Y}|H_0)}{p(\mathcal{Y}|H_1)}$$





DEVELOPMENT

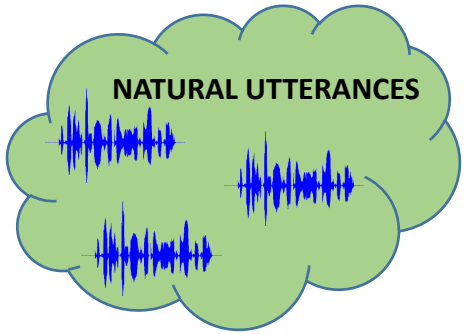
MINIMIZE DETECTION ERROR RATE
ON KNOWN SAMPLES

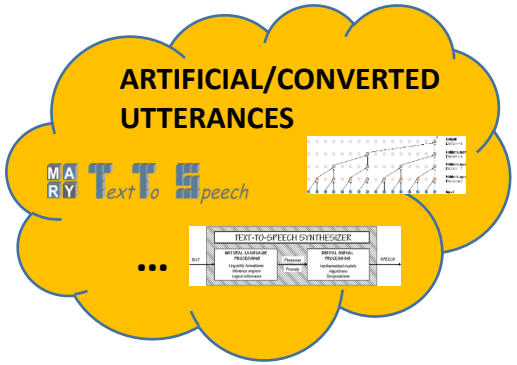
EVALUATION

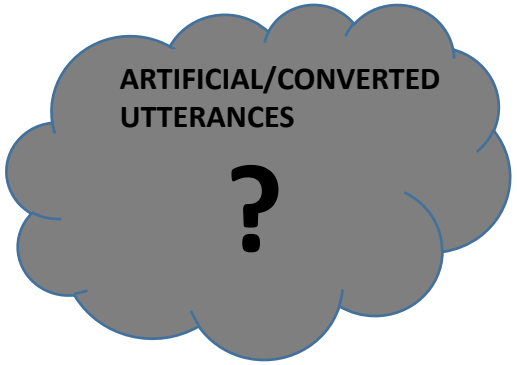
NO OPTIMIZATION ON EVAL DATA !
The higher the error rate, the better the VC system

1. OPTIMIZE COUNTERMEASURE

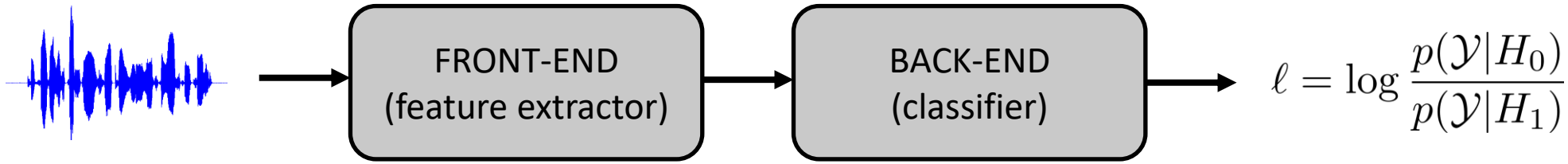
2. ONE-SHOT EXECUTION OF COUNTERMEASURE







Spoofing countermeasure



CONSTANT-Q CEPSTRAL COEFFICIENTS (CQCCs)

Massimiliano Todisco, Hector Delgado and Nicholas Evans: ”A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients”, *Proc. Odyssey 2016*

Implementation: ASVspoof 2017 challenge baseline
http://www.asvspoof.org/data2017/baseline_CM.zip

GAUSSIAN MIXTURE MODEL (GMM)

C. Hanilçi, T. Kinnunen, M. Sahidullah, A. Sizov, ”Classifiers for synthetic speech detection: a comparison”, *Proc. Interspeech 2015,* pp. 2057--2061, Dresden, Germany, September 2015

M. Sahidullah, T. Kinnunen, C. Hanilçi, ”A comparison of features for synthetic speech detection”, *Proc. Interspeech 2015,* pp. 2087--2091, Dresden, Germany, September 2015

Implementation: MSR Identity Toolkit
Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck, “MSR identity toolbox: A MATLAB toolbox for speaker recognition research (v 1.0)”
https://www.microsoft.com/en-us/download/details.aspx?id=52279

Development data

	ASVspoof'15 TRAIN	ASVspoof'15 DEV	ASVspoof'15 EVAL	VCC'16 submissions	VCC'18 Baseline
Attacks	TTS, VC	TTS, VC	TTS, VC	VC	VC
Waveform generation	STRAIGHT, MLSA	STRAIGHT	STRAIGHT, diphone conc.	STRAIGHT, LPC, Ahocoder	Waveform Filtering
# Attacks	5	10	10	18	1
# Humans	25	35	46	10	12
# Spoof trials	12,625	49,875	184,000	24,300	2,240
# Human trials	3,750	3,497	9,404	535	1,139

- STATE-OF-THE-ART IN 2015
- LARGE AMOUNT OF DATA
- HOMOGENOUS WAVEFORM MODELING TECHNIQUES

- STATE-OF-THE-ART IN 2016
- LESS DATA
- MORE VARIED METHODS

BASELINE SYSTEM PROVIDED TO VCC'18 PARTICIPANTS

Dataset

Features: 29 constant-Q cepstral coefficients + energy + Δ + Δ^2
Classifier: Gaussian mixture model (32 Gaussians)

EQUAL ERROR RATE (EER, %) OF HUMAN-vs-SPOOF DISCRIMINATION
THE LOWER, THE BETTER

WITHIN-DATABASE
(sanity check)

TRAINING DATA	TEST DATA	EER (%)
ASVspoof'15 (train)	ASVspoof'15 (dev)	0.38
ASVspoof'15 (train)	ASVspoof'15 (eval)	1.83

ACROSS-DATABASE

TRAINING DATA	TEST DATA	EER (%)
ASVspoof'15 (train)	VCC'16	35.04
ASVspoof'15 (train+dev+eval)	VCC'16	34.86
VCC'16	ASVspoof'15 (dev)	26.18
VCC'16	ASVspoof'15 (eval)	21.48

Countermeasure tuning

CQCC CONFIGURATION

32 Gaussians, training: VCC 2016

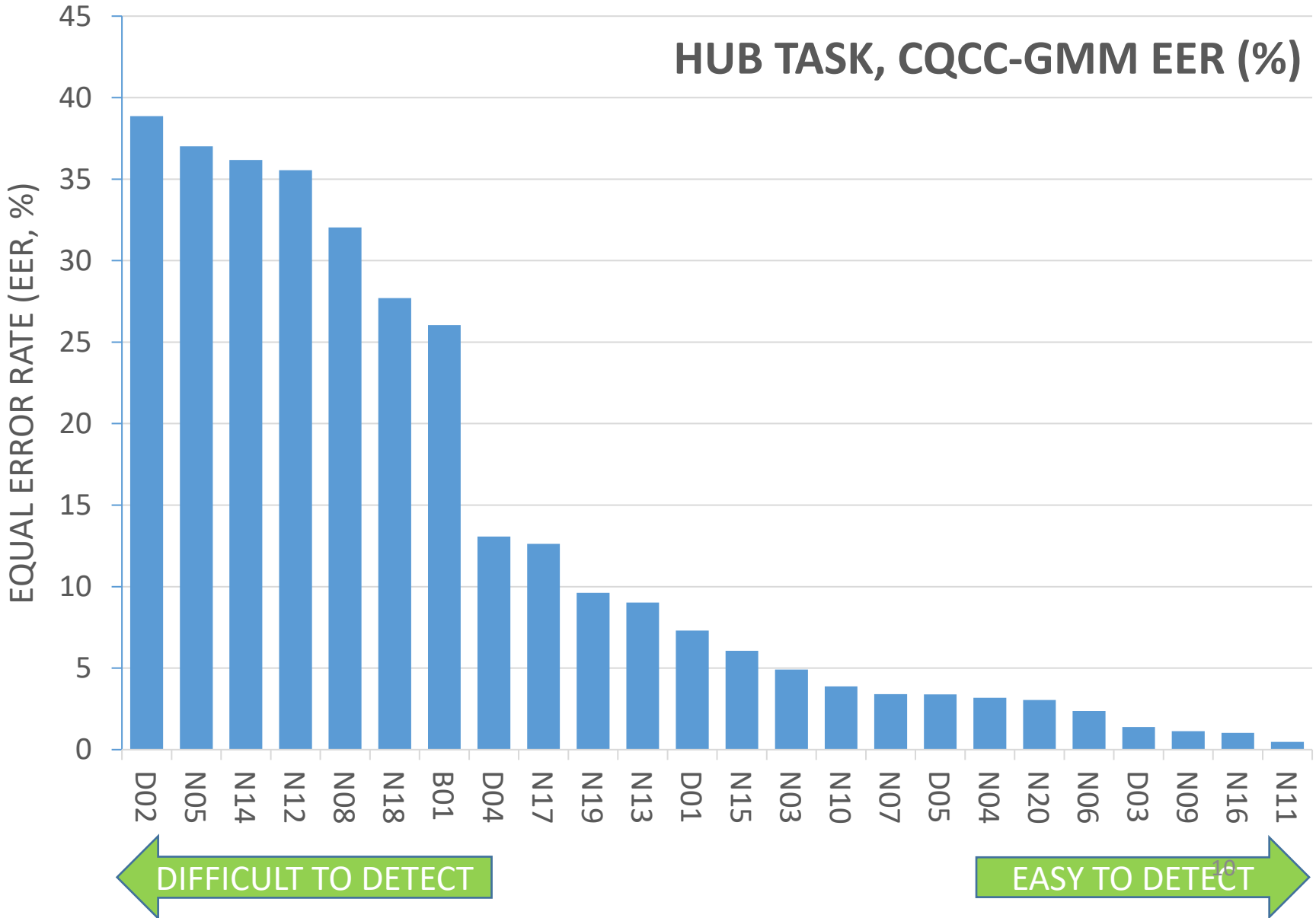
FRONT-END	EVALUATE ON: ASVspoof'15 (dev)	EVALUATE ON: VCC'18 baseline
Static	46.14	18.41
Δ	12.11	24.85
Δ^2	13.58	22.89
Δ, Δ^2	7.73	22.08
Static, Δ	39.73	15.45
Static, Δ^2	34.46	13.83
Static, Δ, Δ^2	32.09	14.00
Static, c_0, Δ, Δ^2	26.18	14.13

FIX THIS CONFIGURATION,
SET #GAUSSIANS = 2048

One-shot scoring – VCC HUB (parallel VC training)

CORRECTED RESULTS (bug found in post-evaluation)

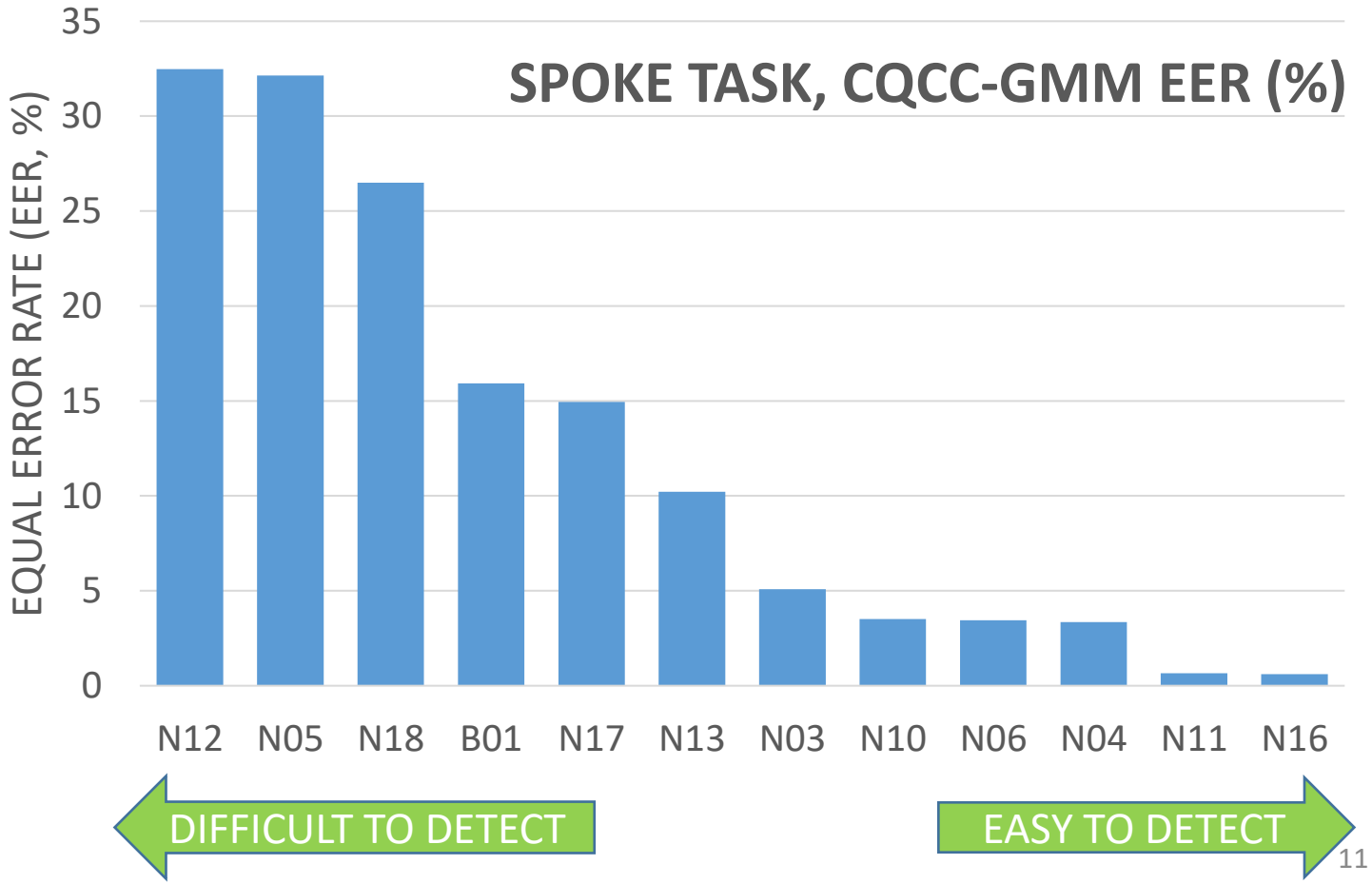
System	Waveform generation	EER (%)
D02	WORLD	38.9
N05	SuperVP	37.0
N14	Waveform filtering	36.2
N12	Waveform filtering	35.6
N08	Waveform filtering	32.0
N18	Griffin-Lim	27.7
B01	Waveform filtering	26.0
D04	WORLD	13.1
N17	Wavenet	12.6
N19	WORLD	9.6
N13	WORLD	9.0
D01	WORLD direct wave modif.	7.3
N15	WORLD	6.1
N03	(*unknown*)	4.9
N10	Wavenet	3.9
N07	WORLD	3.4
D05	WORLD	3.4
N04	WORLD	3.2
N20	WORLD	3.0
N06	WORLD	2.4
D03	STRAIGHT	1.4
N09	Ahocoder	1.1
N16	Ahocoder	1.0
N11	STRAIGHT	0.5

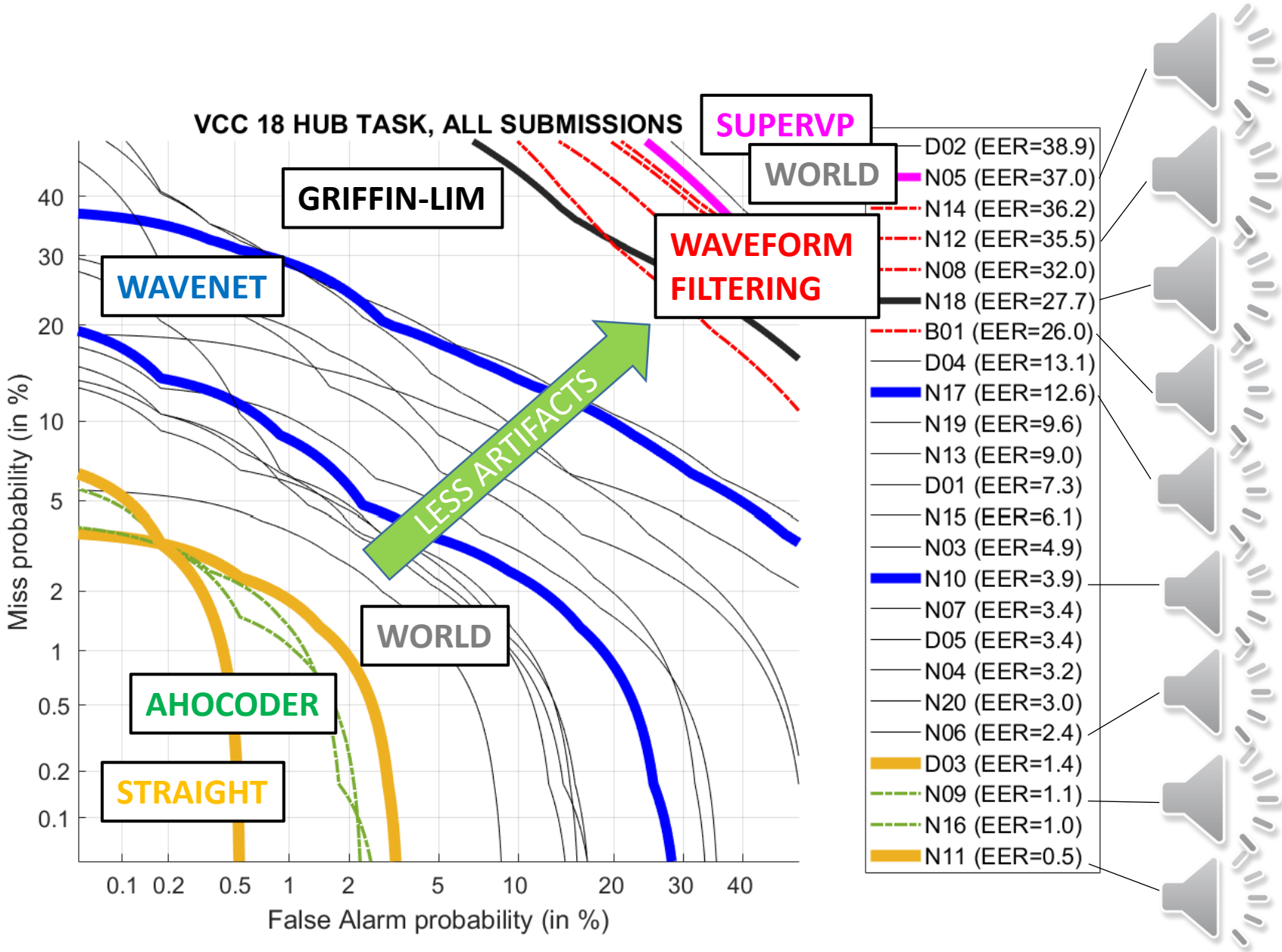


One-shot scoring – VCC SPOKE (nonparallel VC training)

CORRECTED RESULTS (bug found in post-evaluation)

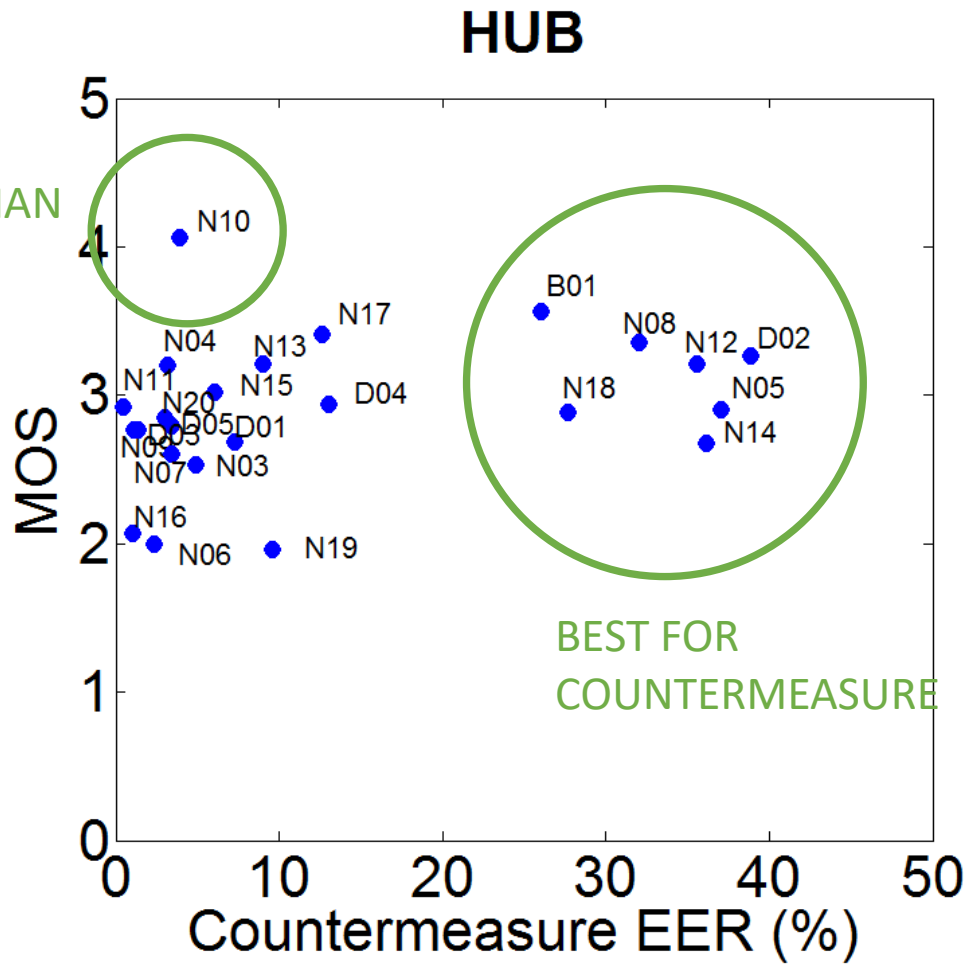
System	Waveform generation	EER (%)
N12	Waveform filtering	32.5
N05	SuperVP	32.1
N18	Griffin-Lim	26.5
B01	Waveform filtering	15.9
N17	Wavenet	14.9
N13	WORLD	10.2
N03	(unknown)	5.1
N10	Wavenet	3.5
N06	WORLD	3.4
N04	WORLD	3.4
N11	STRAIGHT	0.7
N16	Ahocoder	0.6



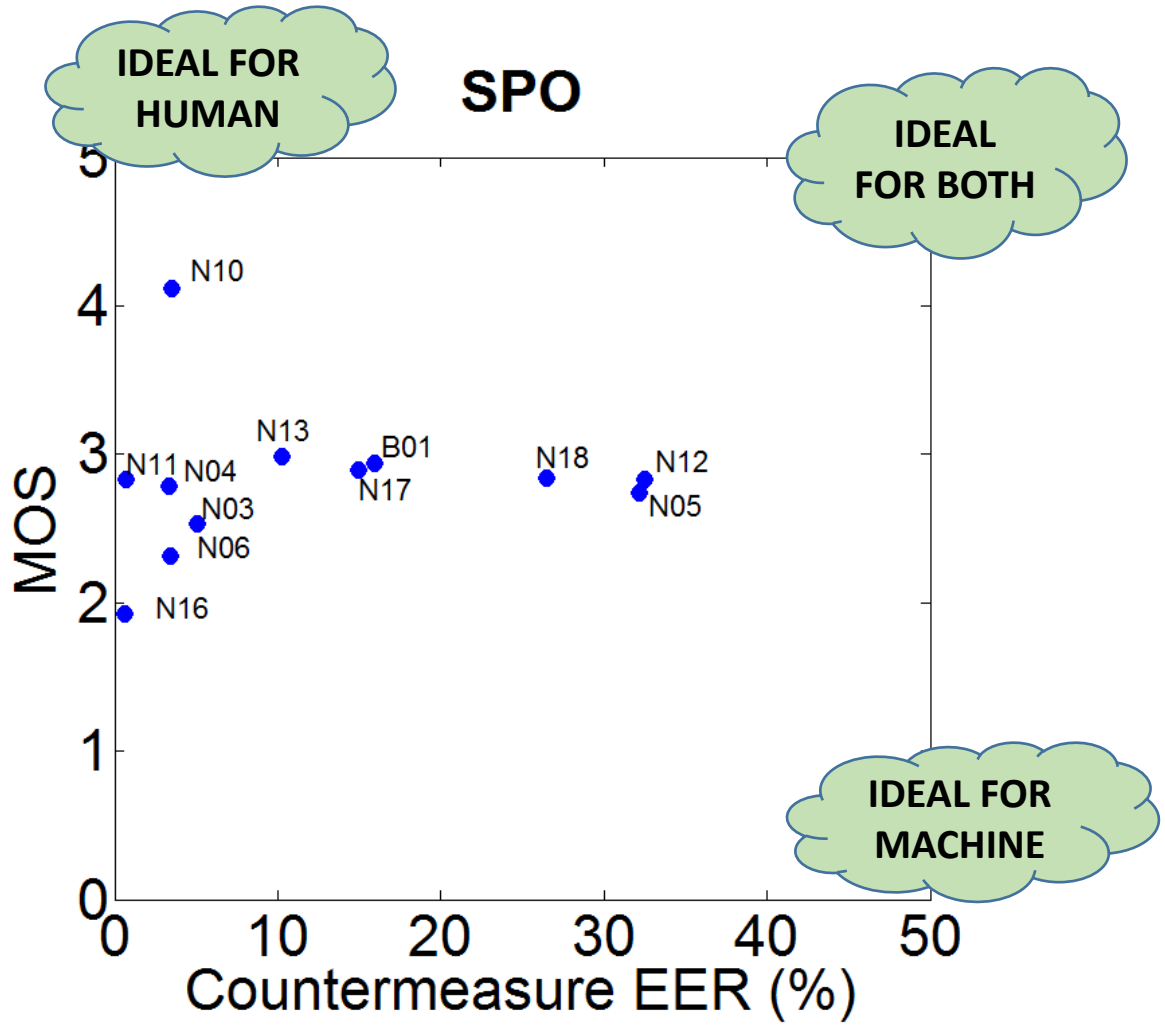

VCC 18 HUB TASK, ALL SUBMISSIONS
SUPERVP
WORLD
GRIFFIN-LIM
WAVEFORM FILTERING
WAVENET
LESS ARTIFACTS
WORLD
AHOCODER
STRAIGHT
Miss probability (in %)
False Alarm probability (in %)
40
30
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 30 40
D02 (EER=38.9)
N05 (EER=37.0)
N14 (EER=36.2)
N12 (EER=35.5)
N08 (EER=32.0)
N18 (EER=27.7)
B01 (EER=26.0)
D04 (EER=13.1)
N17 (EER=12.6)
N19 (EER=9.6)
N13 (EER=9.0)
D01 (EER=7.3)
N15 (EER=6.1)
N03 (EER=4.9)
N10 (EER=3.9)
N07 (EER=3.4)
D05 (EER=3.4)
N04 (EER=3.2)
N20 (EER=3.0)
N06 (EER=2.4)
D03 (EER=1.4)
N09 (EER=1.1)
N16 (EER=1.0)
N11 (EER=0.5)

Objective and subjective quality are complementary

N10 BEST FOR HUMAN

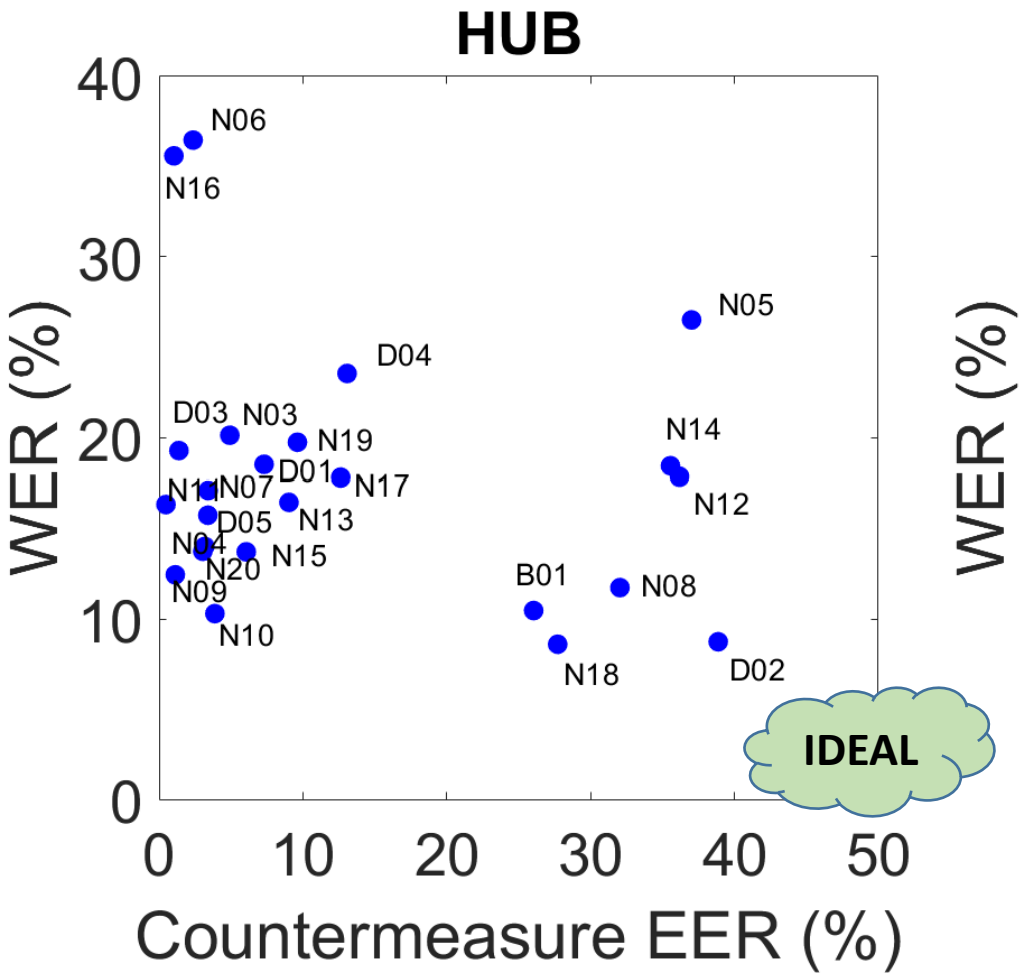


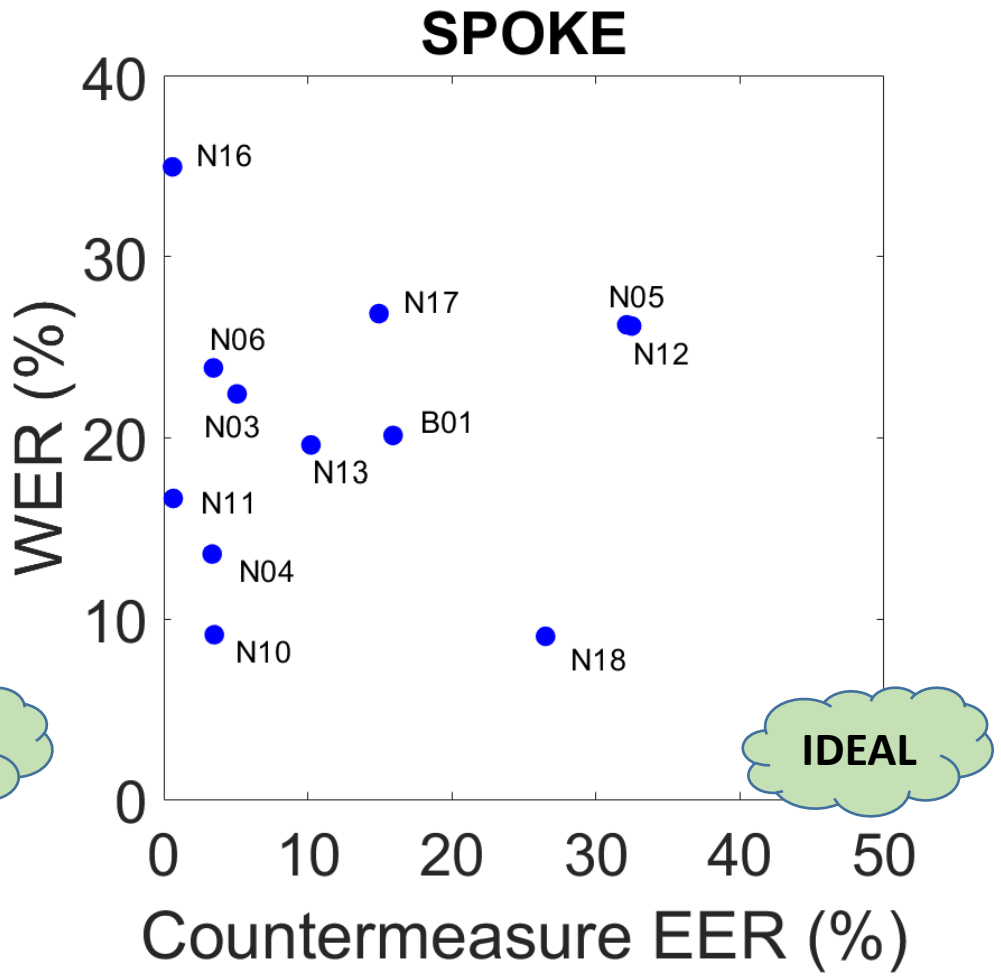
r = 0.3766



r = 0.3725

Machine-only ”listener”: CQCC EER vs. WER





r = -0.2309

r = -0.0745

Take-home messages

FOR ASV COMMUNITY

- VC technology has taken a substantial leap since 2015 ➔ EERs up to ~40%
- Need to improve cross-dataset performance of countermeasures

FOR VC COMMUNITY

- Weak correlation between MOS and countermeasure EER
- Spoofing countermeasure as a new objective evaluation tool ?