

Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions



Rohan Kumar Das¹, Tomi Kinnunen², Wen-Chin Huang³, Zhenhua Ling⁴,
Junichi Yamagishi⁵, Yi Zhao⁵, Xiaohai Tian¹ and Tomoki Toda³

¹National University of Singapore, Singapore

²University of Eastern Finland, Joensuu, Finland ³Nagoya University, Japan

⁴University of Science and Technology of China, China

⁵National Institute of Informatics, Japan

Investigation

- Can the **objective assessment metrics** predict human judgements on **naturalness** and **speaker similarity**?
- Which **voice conversion** (VC) technology has the **highest spoofing risk** for automatic speaker verification (ASV) and spoofing countermeasure (CM)?

Need of Objective Assessments

- Complementary to listening tests
- Less time consuming than listening tests
- Cost effective than large crowd sourcing listening tests



+



=



ASV Vulnerability to Spoofing Attacks

banking
technology

Get your copy of the 2017 Digital Sales Readiness Matrix.

CLICK HERE



HOME NEWS SIBOS MAGAZINES AWARDS RESOURCES EVENTS JOBS MORE

Search

GO

Home » Region » UK » Twins win in HSBC voice tricking sting

Twins win in HSBC voice tricking sting

19 May, 2017 Written by Antony Peyton



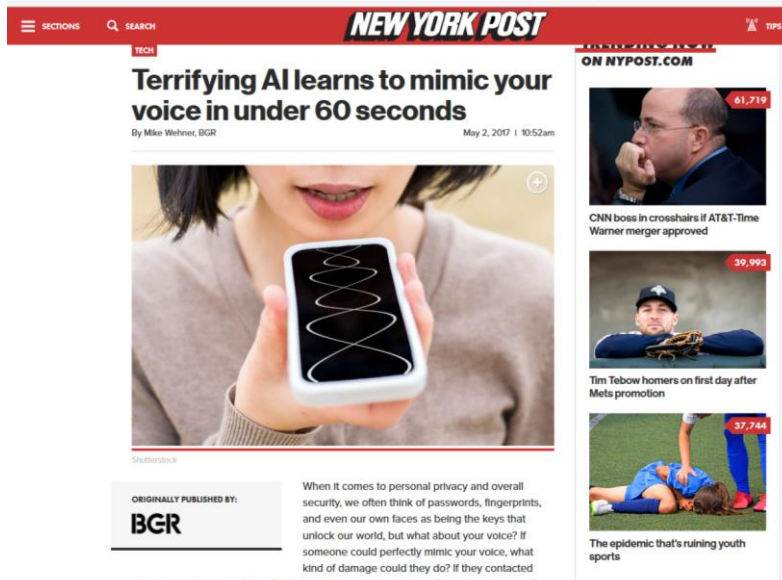
Print



Email

SIGN UP TO OUR DAILY NEWS DIGEST

Receive FREE Banking Technology news alerts straight to your inbox [Sign me up](#)



Spoofing Countermeasures

Nuance deploys AI biometric security tools

19 May 2017 15:51 GMT



[Jump to comments](#)



Biometrics firm Nuance, which has focused on voice recognition, has announced a new multi-modal suite of biometric security solutions, driven by artificial intelligence (AI).

The new suite features facial and behavioural biometrics, as well as voice, with the company saying that these combine to provide advanced protection against fraud

Nuance has said that deep neural networks (DNN) are being used in the new solution along side advanced algorithms to detect “synthetic speech attacks”.

“By combining a range of physical, behavioural, and digital characteristics to provide secure authentication and more accurately detect fraud across multiple channels - from the phone to the Web, mobile apps and more - Nuance’s new Security Suite allows enterprises to attack fraud head-on, while at the same time offering an improved customer experience”, wrote the firm.

In particular, the firm notes improved synthetic speech detection

Automatic Speaker Verification

Spoofing And Countermeasures Challenge

ASVspoo Challenge

<https://www.asvspoof.org>

Research on Spoofing
Countermeasures for attacks
derived using

Voice conversion (VC)
Text-to-speech (TTS)
Replay speech

Objective evaluation techniques

Automatic speaker verification (ASV) - speaker similarity

- x-vector based speaker embedding [1]
- PLDA for scoring & cosine similarity of speaker embeddings

Spoofing countermeasure (CM) - real vs. fake assessment

- Light CNN system [2] with LFCC features
- Trained on ASVspoof 2019 logical access corpus training set

Automatic mean opinion score (MOSNet) - quality

- CNN-LSTM with magnitude spectrum as input, following [3]
- Training data: VCC2018/ASVspoof2019

Automatic speech recognition (ASR) - intelligibility

- A prototype system by iFlytek: Seq2seq with attention [4]
- 10,000-hrs recordings for AM / GB-level texts for LM modeling

[1] <https://kaldi-asr.org/models/m7>

[2] G. Lavrentyva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlos, "STC antispoofing systems for the ASVspoof2019 challenge," in Interspeech 2019, 2019, pp. 1033– 1037.

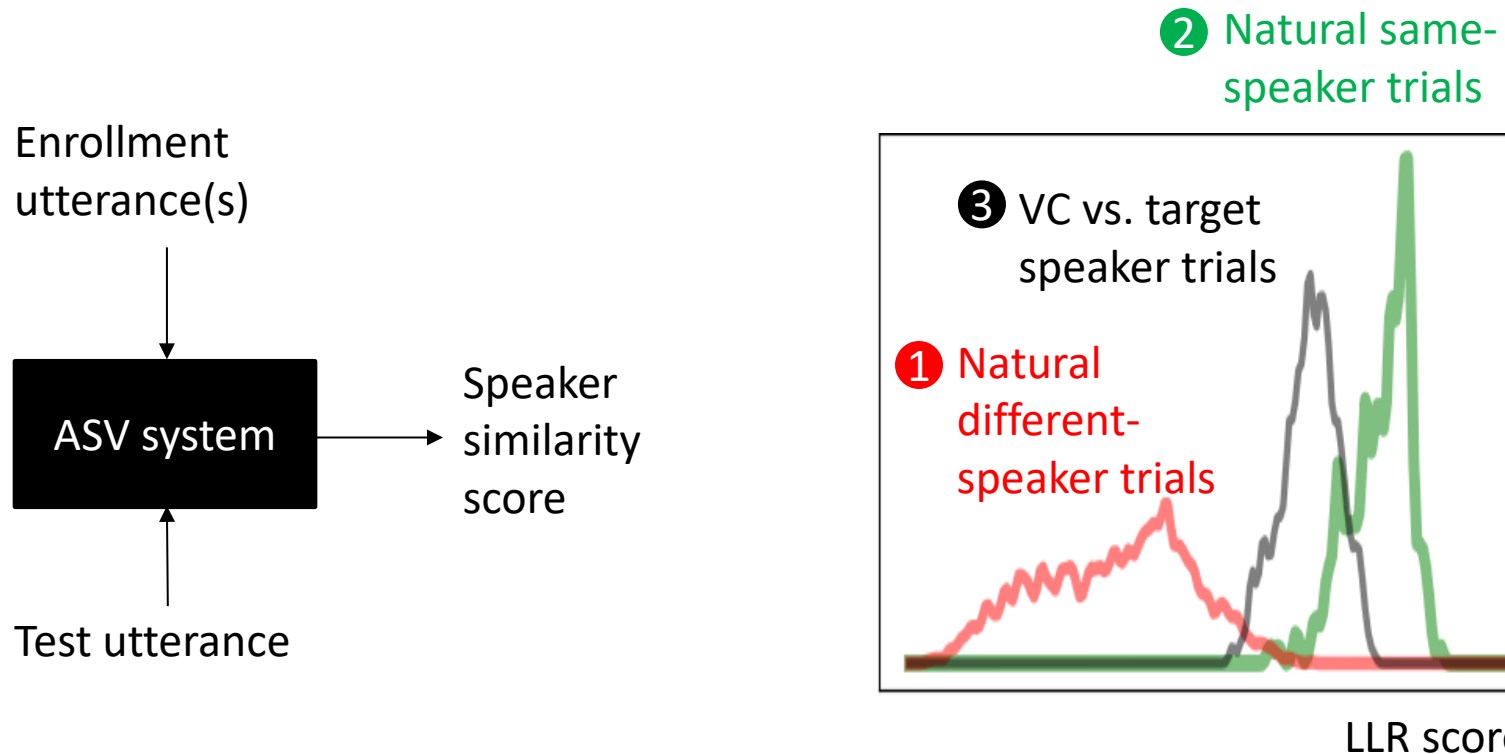
[3] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, H.-M. Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in Proc. Interspeech 2019, pp. 1541-1545

[4] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016, 2016, pp. 4945–4949.

Objective Evaluation Results

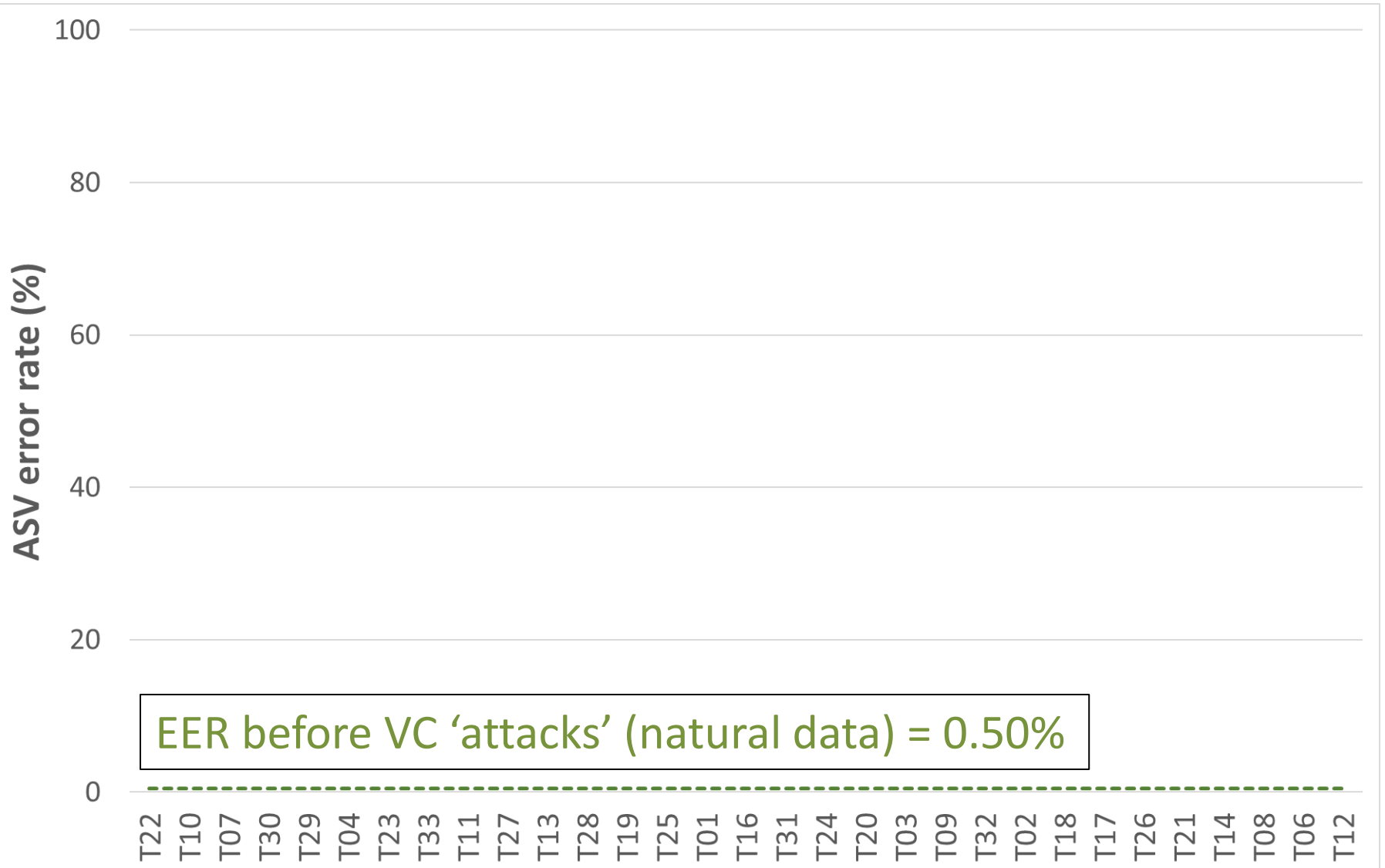
ASV - objective speaker similarity

Kaldi VoxCeleb x-vector PLDA recipe



System-level VC success metric =
overlap of ② & ③ measured by Equal Error Rate (EER)

ASV – Task 1



ASV – Task 1

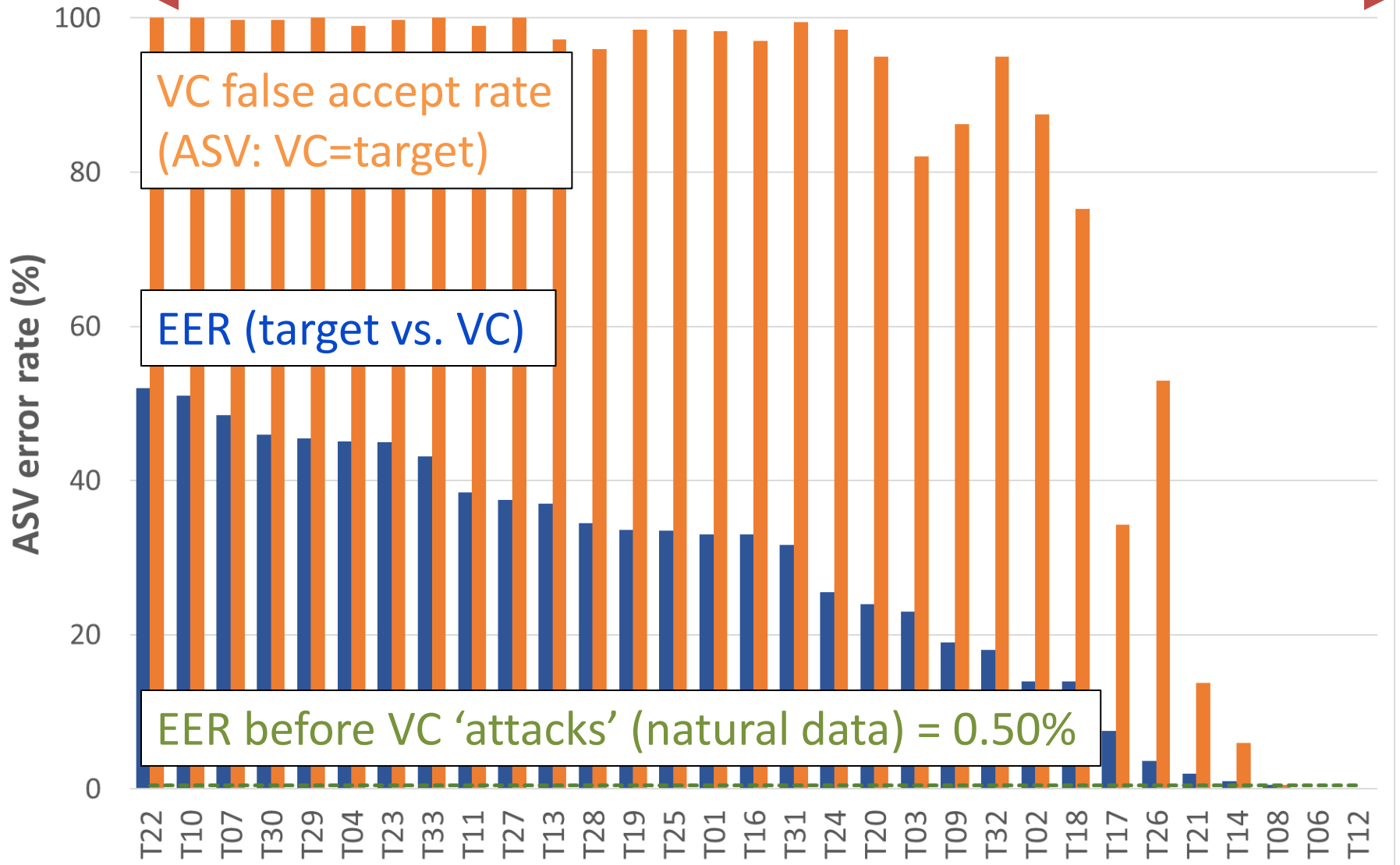
High speaker similarity

Low speaker similarity

VC false accept rate
(ASV: VC=target)

EER (target vs. VC)

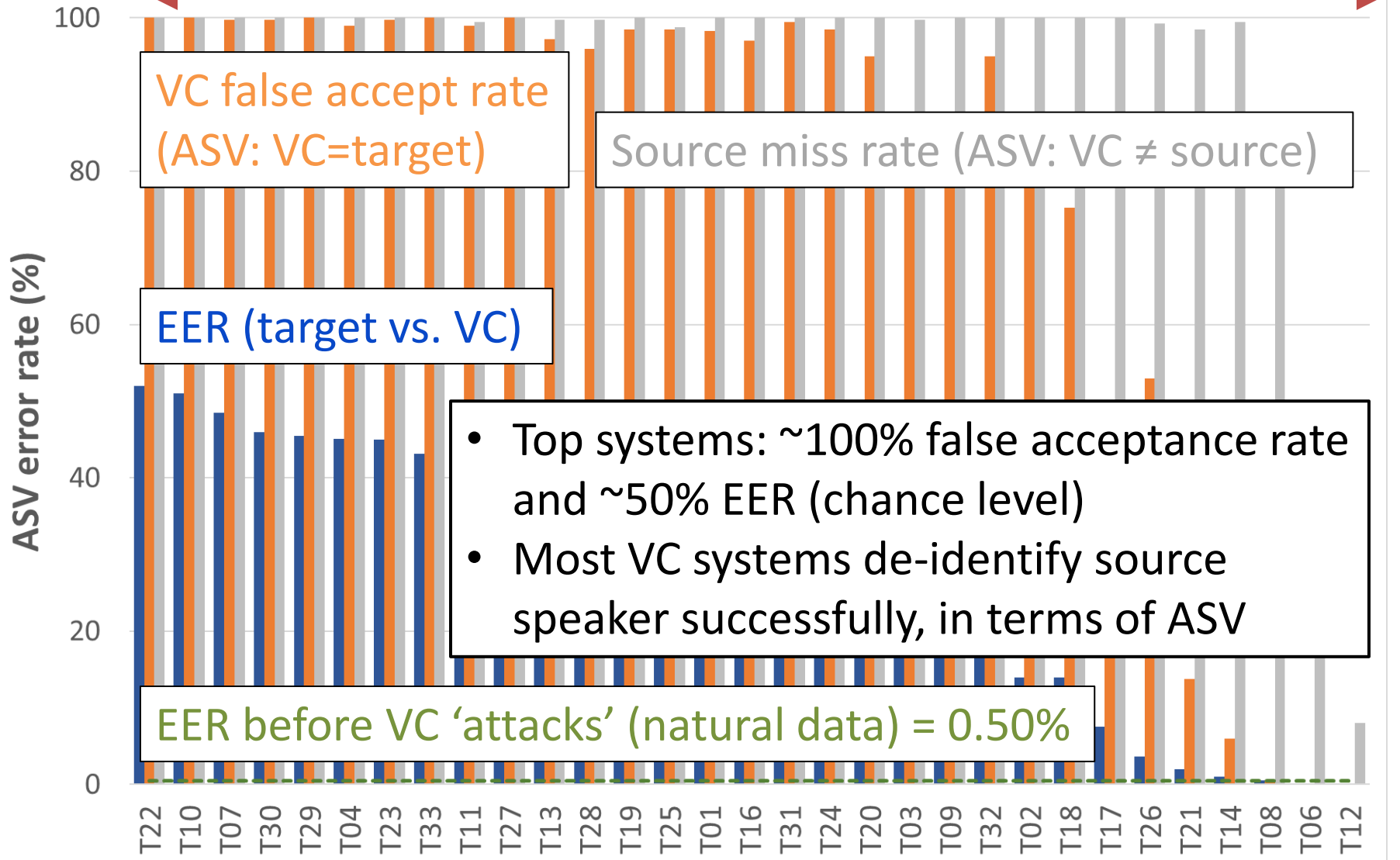
EER before VC 'attacks' (natural data) = 0.50%



ASV – Task 1

High speaker similarity

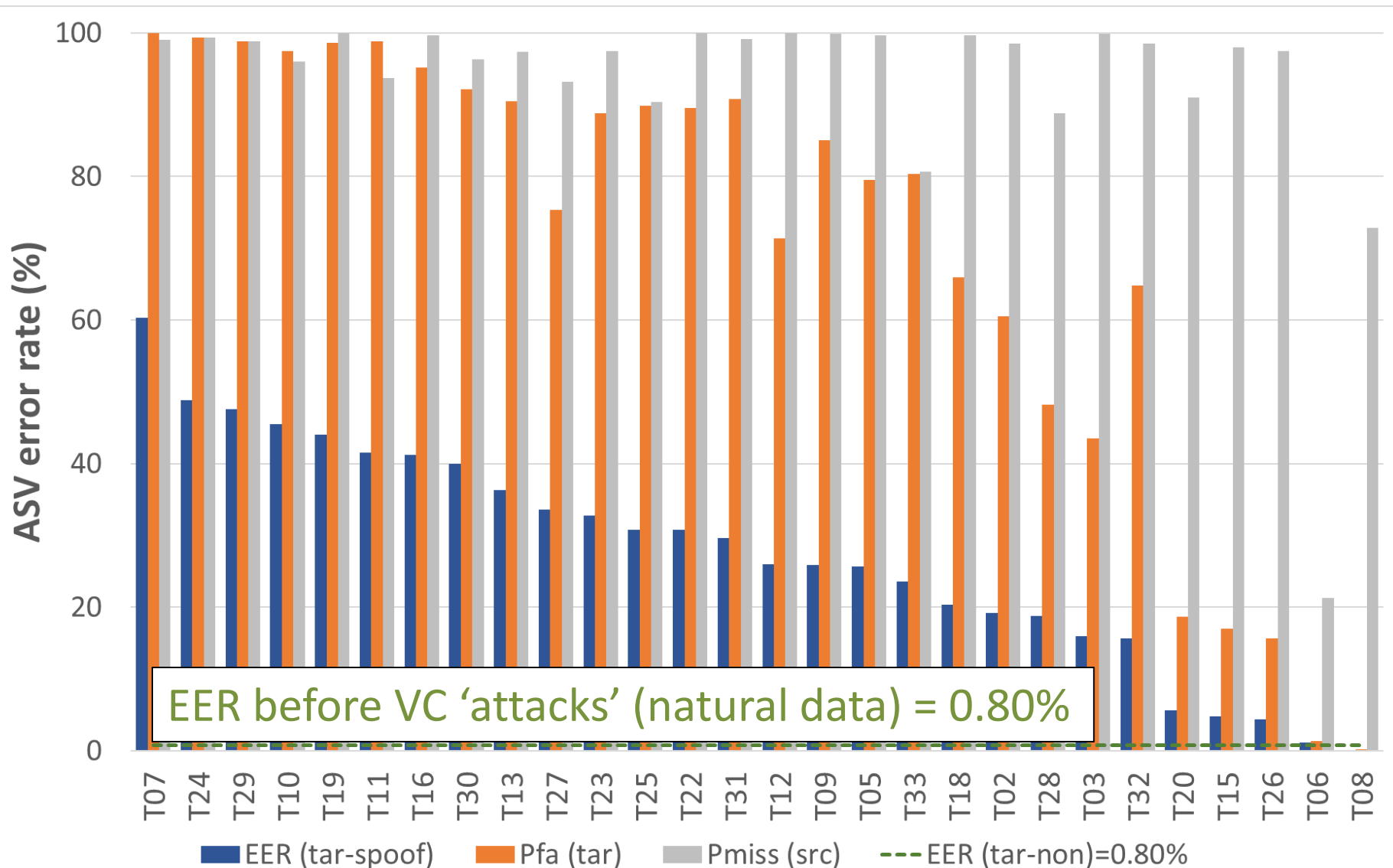
Low speaker similarity



ASV, Task 2 – similar trends

High speaker similarity

Low speaker similarity

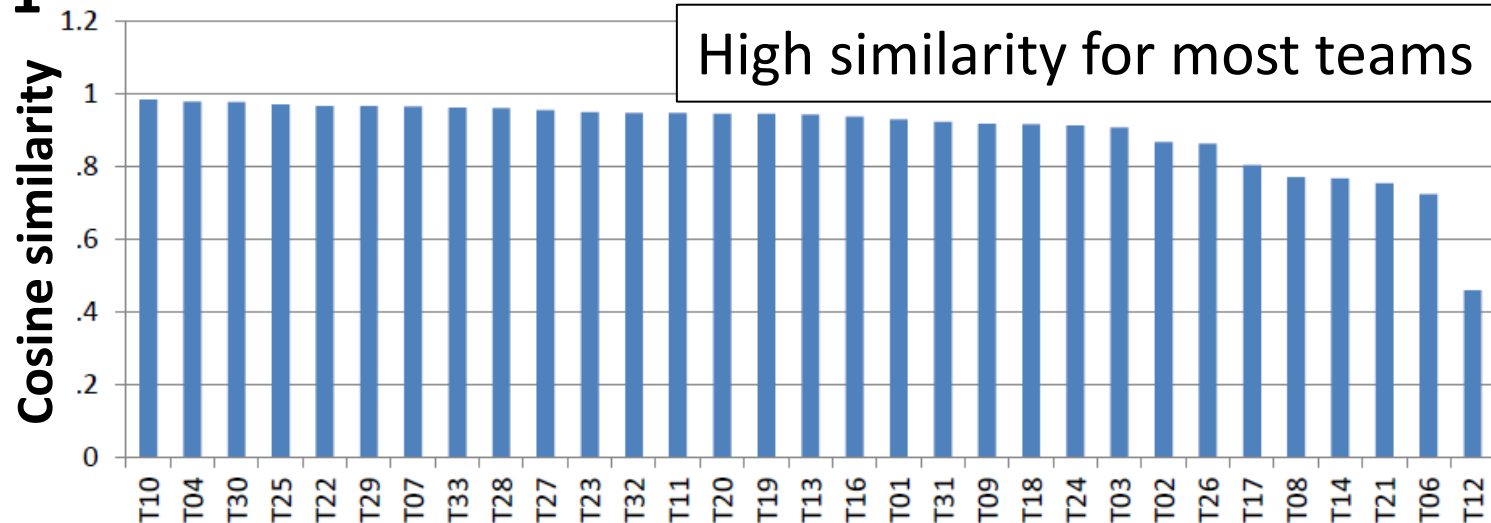


Cosine similarity

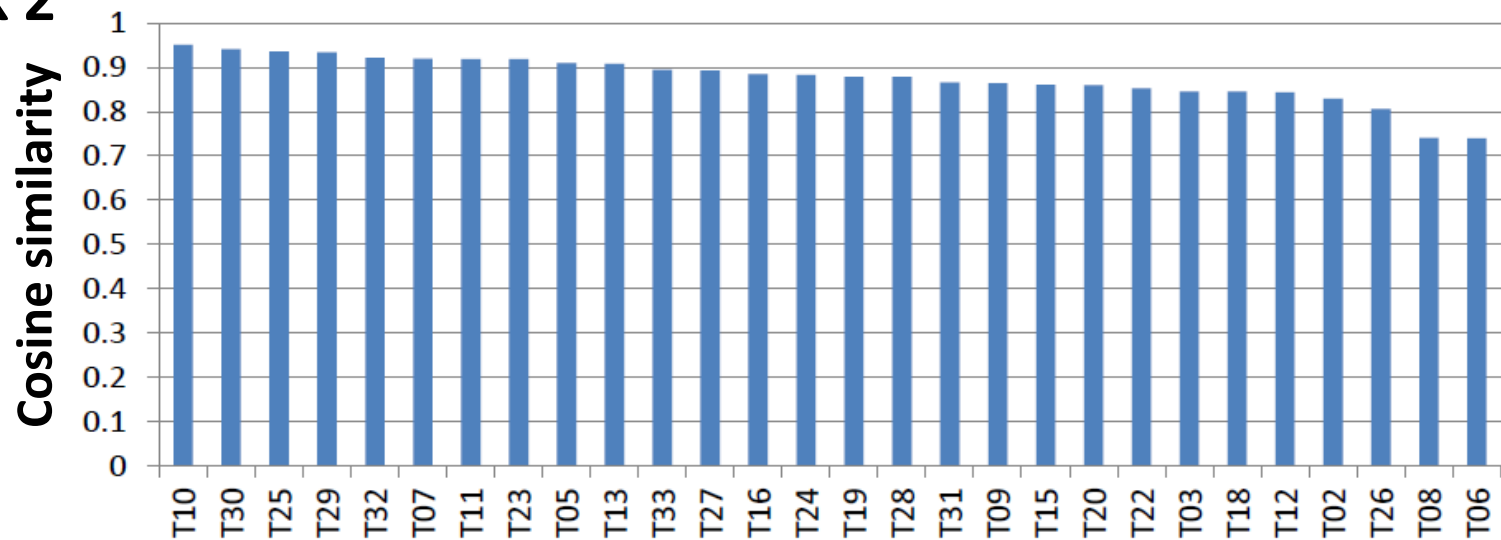
High speaker similarity

Low speaker similarity

TASK 1



TASK 2



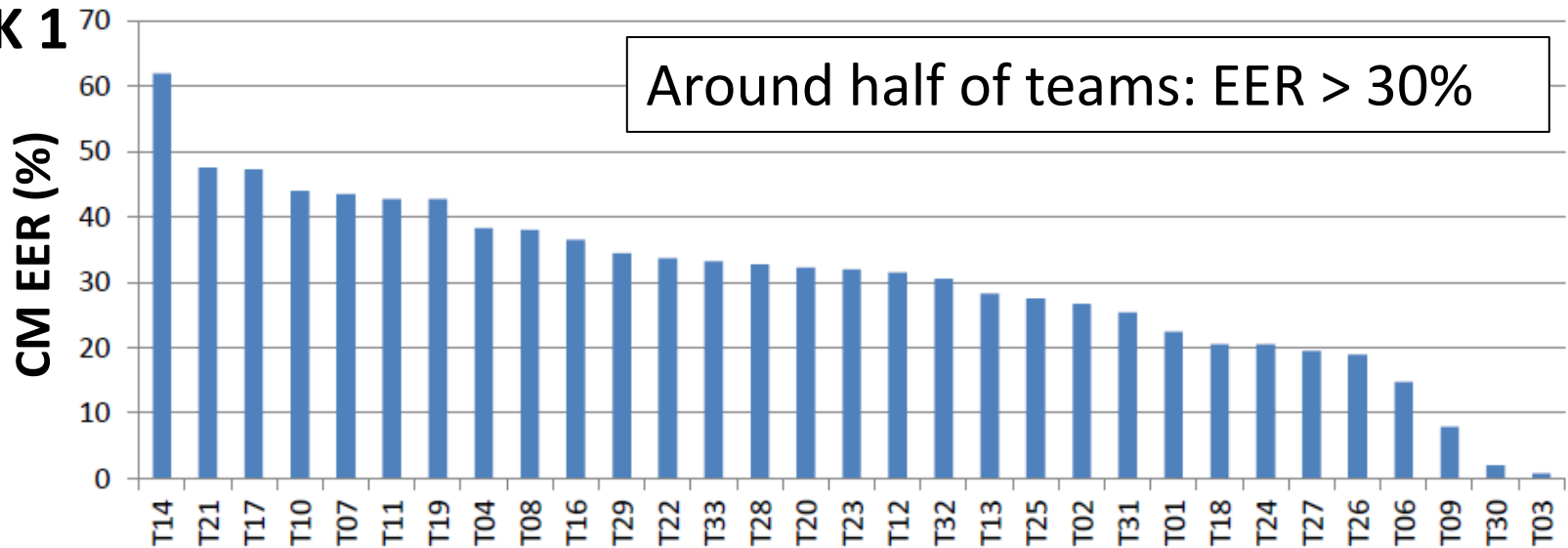
Spoofing countermeasure

Less artifacts

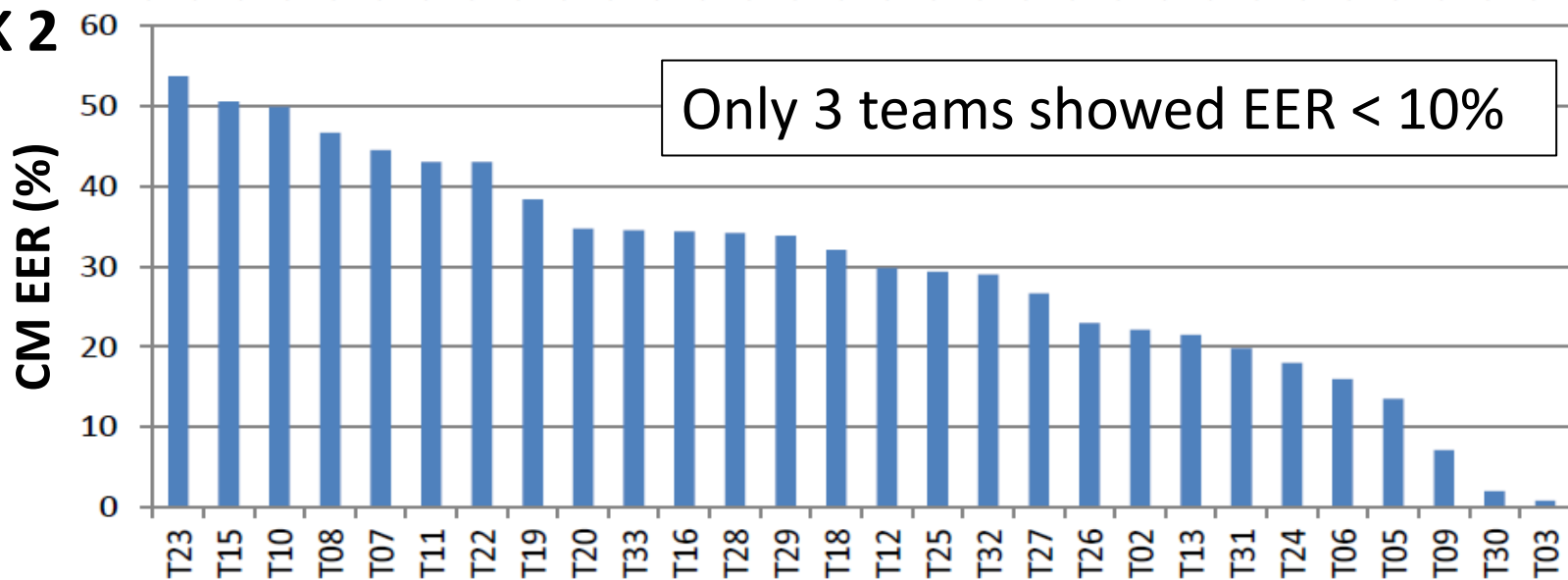
More artifacts



TASK 1



TASK 2

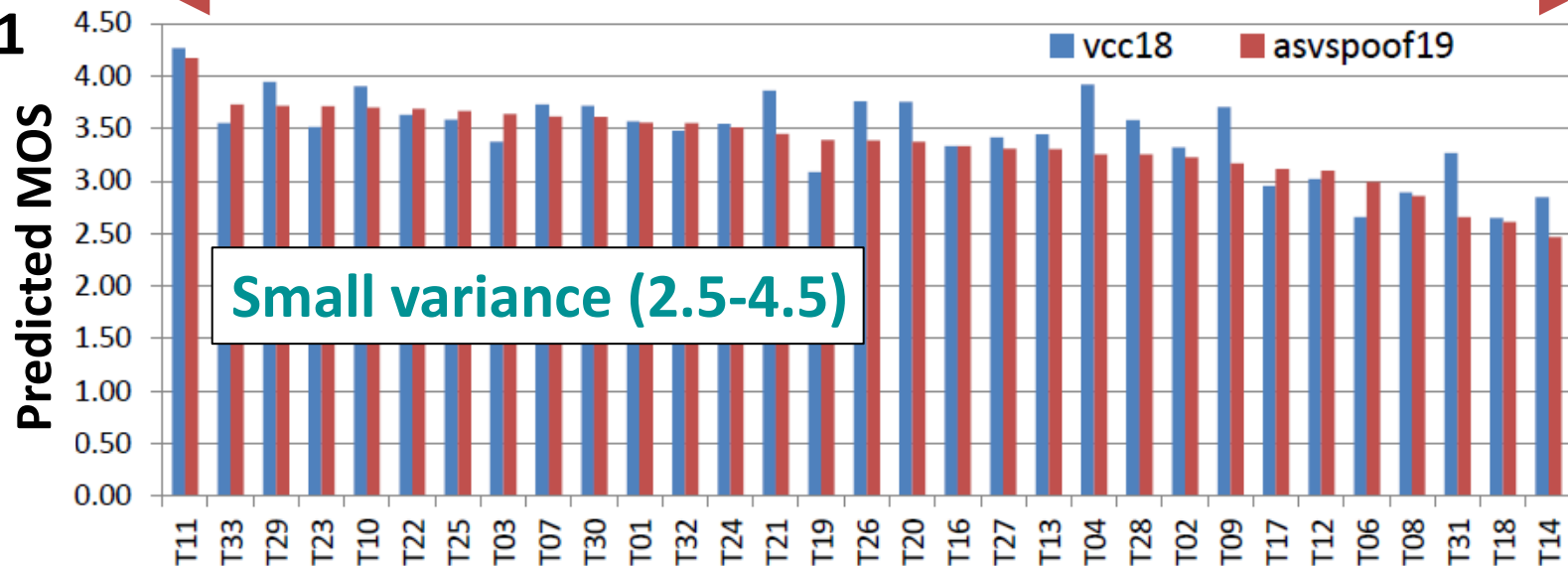


MOSNet

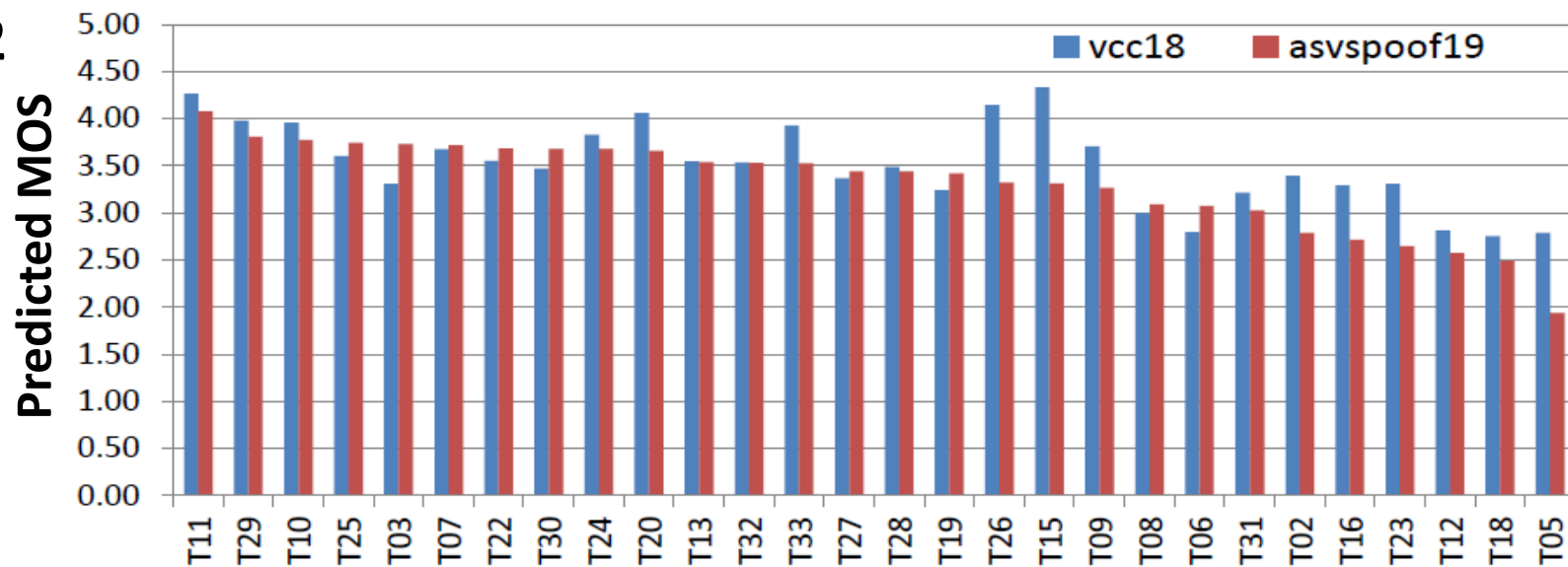
Higher quality

Lower quality

TASK 1



TASK 2



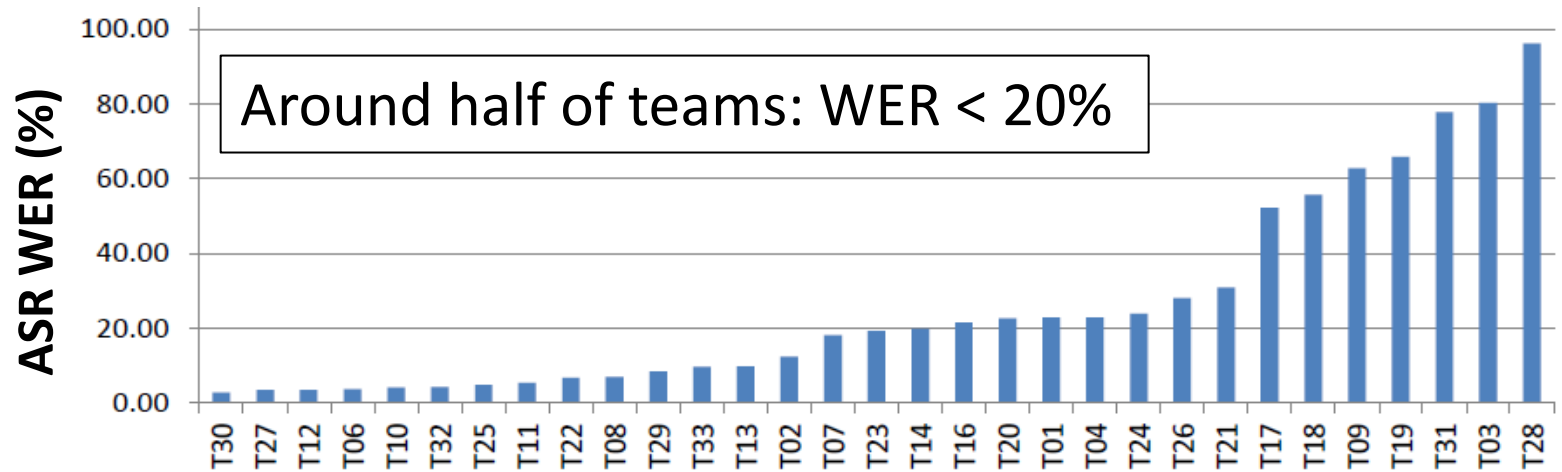
Automatic speech recognition

High intelligibility

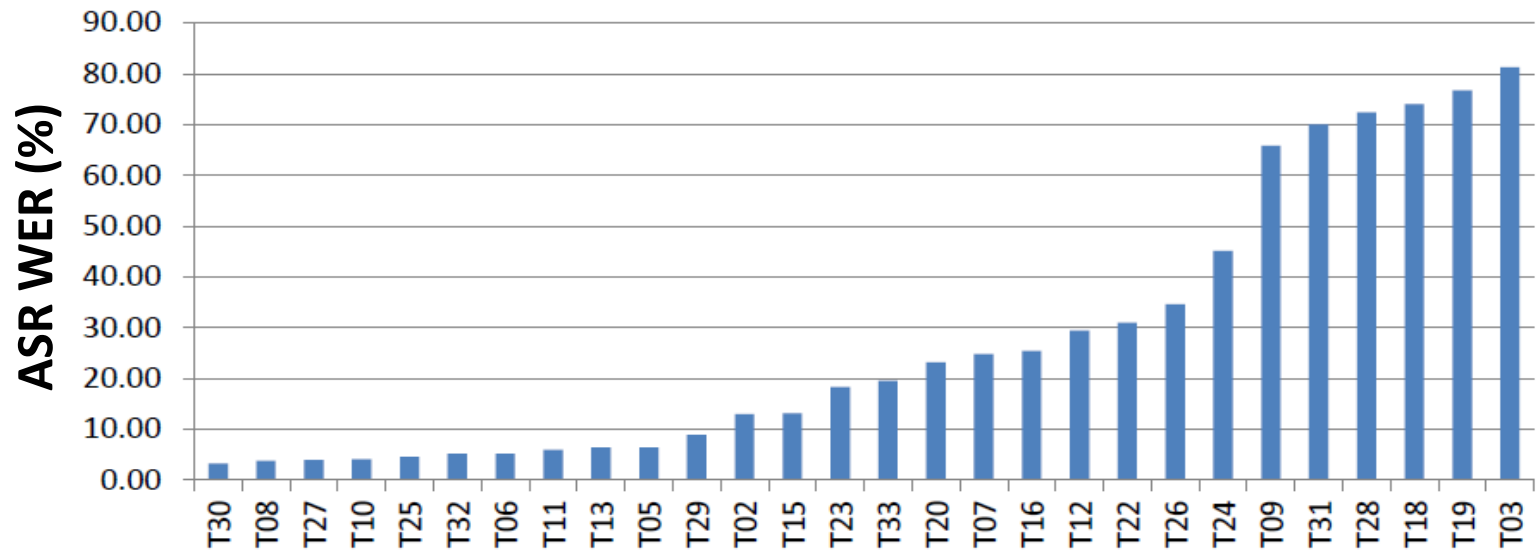
Low intelligibility















TASK 1



TASK 2



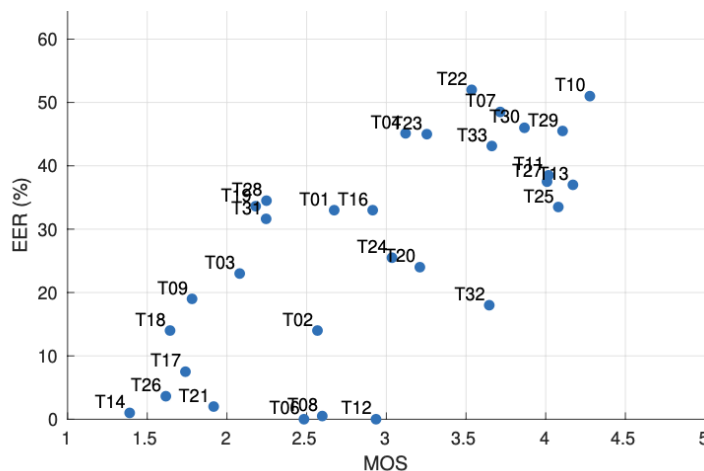
Audio samples

Metric	Description	Score	Src	Trg	VC
ASV	Task 1, Team 22 SEM2-TEM1-E30012	LLR=53.91285			
CM	Task 2, Team 22 SEF1-TGM1-E30009	CM score = 0.9984			
MOSNet (asvspoof19)	Task 1, Team 14 SEM1-TEF1-E30013	MOS = 2.47			
ASR	Task 2, Team 18 SEM2-TMM1-E30010	WER = 91.67			

Correlation with subjective results

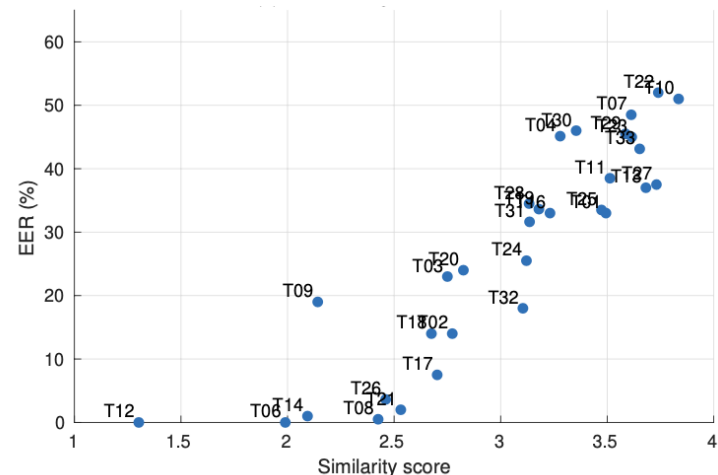
Correlation with Subjective Test Results

- Can the metrics predict human judgements?
- Analysis 1: Draw scatter plots (appendix of paper)



ASV EER (%), task 1 MOS

Lower correlation?




ASV EER (%), task 1 SIM

Higher correlation?

- Analysis 2: **Calculate Pearson correlation coefficients**


Individual Pearson correlation coefficients



Subjective score (ENG)	ASV EER (%)	ASV Pfa (%)	Cosine distance	Countermeasure EER (%)	MOSNet (vcc18)	MOSNet (asvspoof19)	ASR WER (%)
Task 1 MOS	0.70 ($p < 0.01$)	0.53 ($p < 0.01$)	0.42 ($p > 0.01$)	0.00 ($p > 0.01$)	0.52 ($p < 0.01$)	0.66 ($p < 0.01$)	-0.65 ($p < 0.01$)
Task 1 SIM	0.89 ($p < 0.01$)	0.82 ($p < 0.01$)	0.85 ($p < 0.01$)	0.07 ($p > 0.01$)	0.54 ($p < 0.01$)	0.61 ($p < 0.01$)	-0.18 ($p > 0.01$)
Task 2 MOS	0.34 ($p > 0.01$)	0.26 ($p > 0.01$)	0.59 ($p < 0.01$)	0.27 ($p > 0.01$)	0.43 ($p > 0.01$)	0.58 ($p < 0.01$)	-0.73 ($p < 0.01$)
Task 2 SIM	0.90 ($p < 0.01$)	0.86 ($p < 0.01$)	0.82 ($p < 0.01$)	0.19 ($p > 0.01$)	0.23 ($p > 0.01$)	0.32 ($p > 0.01$)	-0.14 ($p > 0.01$)

- Metrics with moderate (>0.5) coefficients for **quality**:
 Task 1: **ASV (EER, Pfa)**, **MOSNet (vcc18, asvspoof19)**, **ASR WER**
 Task 2: **cosine distance**, **MOSNet (asvspoof19)**, **ASR WER**
- Why do ASV and cosine distance show high correlation?
- Human Judgements on quality and similarity are **not independent!**

Individual Pearson correlation coefficients



Subjective score (ENG)	ASV EER (%)	ASV Pfa (%)	Cosine distance	Countermeasure EER (%)	MOSNet (vcc18)	MOSNet (asvspoof19)	ASR WER (%)
Task 1 MOS	0.70 ($p < 0.01$)	0.53 ($p < 0.01$)	0.42 ($p > 0.01$)	0.00 ($p > 0.01$)	0.52 ($p < 0.01$)	0.66 ($p < 0.01$)	-0.65 ($p < 0.01$)
Task 1 SIM	0.89 ($p < 0.01$)	0.82 ($p < 0.01$)	0.85 ($p < 0.01$)	0.07 ($p > 0.01$)	0.54 ($p < 0.01$)	0.61 ($p < 0.01$)	-0.18 ($p > 0.01$)
Task 2 MOS	0.34 ($p > 0.01$)	0.26 ($p > 0.01$)	0.59 ($p < 0.01$)	0.27 ($p > 0.01$)	0.43 ($p > 0.01$)	0.58 ($p < 0.01$)	- 0.73 ($p < 0.01$)
Task 2 SIM	0.90 ($p < 0.01$)	0.86 ($p < 0.01$)	0.82 ($p < 0.01$)	0.19 ($p > 0.01$)	0.23 ($p > 0.01$)	0.32 ($p > 0.01$)	-0.14 ($p > 0.01$)

- Metrics with moderate (>0.5) coefficients for **similarity**:

Task 1: ASV (EER, Pfa), cosine distance,
MOSNet (vcc18, asvspoof19)

Task 2: ASV (EER, Pfa), cosine distance

- High correlation of MOSNet underpin that human Judgements on quality and similarity are **not independent**.

Prediction of Subjective Evaluation Results by Objective Metrics Combinations

Subjective score (ENG)	Intercept	MOSNet (asvspoof19)	ASR WER (%)	ASV EER (%)	Countermeasure EER (%)	Multiple R-Squared	Adjusted R-squared	Significance F
Task 1 MOS	1.713 ($p=0.054$)	0.258 ($p=0.333$)	-0.021 ($p<0.001$)	0.024 ($p<0.001$)	-0.002 ($p=0.654$)	0.92	0.81	<0.001
Task 1 SIM	1.696 ($p=0.006$)	0.062 ($p=0.722$)	-0.003 ($p=0.146$)	0.026 ($p<0.001$)	0.006 ($p=0.108$)	0.92	0.83	<0.001
Task 2 MOS	0.619 ($p=0.357$)	0.724 ($p=0.001$)	-0.021 ($p<0.001$)	0.014 ($p=0.049$)	0.003 ($p=0.668$)	0.88	0.74	<0.001
Task 2 SIM	1.782 ($p<0.001$)	0.038 ($p=0.617$)	-0.002 ($p=0.254$)	0.026 ($p<0.001$)	0.002 ($p=0.538$)	0.91	0.80	<0.001

- Significant explainable variables for **MOS**:
 - Task 1: **ASV EER, ASR WER**
 - Task 2: **MOSNet (asvspoof19), ASR WER**
- Significant explainable variables for **SIM**:
 - Task 1 & 2: **ASV EER** only
 - ASV EER itself has sufficiently high correlation.
- Overall, consistent with previous analysis.

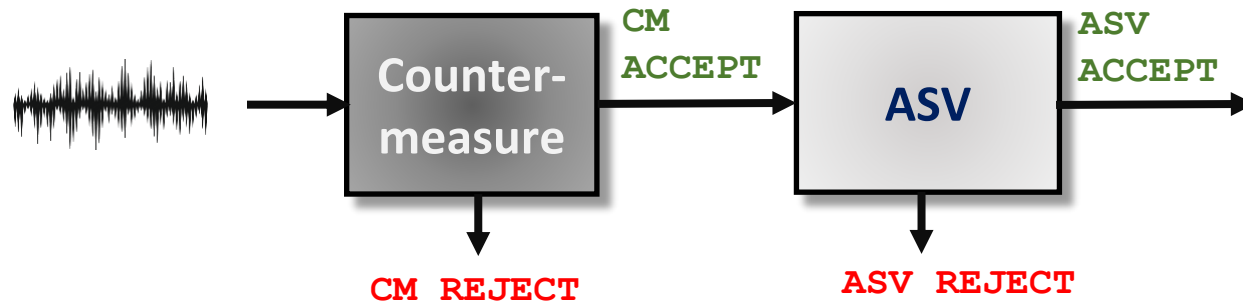
Prediction of Subjective Evaluation Results by Objective Metrics Combinations

Subjective score (ENG)	Intercept	MOSNet (asvspoof19)	ASR WER (%)	ASV EER (%)	Countermeasure EER (%)	Multiple R-Squared	Adjusted R-squared	Significance F
Task 1 MOS	1.713 ($p=0.054$)	0.258 ($p=0.333$)	-0.021 ($p<0.001$)	0.024 ($p<0.001$)	-0.002 ($p=0.654$)	0.92	0.81	<0.001
Task 1 SIM	1.696 ($p=0.006$)	0.062 ($p=0.722$)	-0.003 ($p=0.146$)	0.026 ($p<0.001$)	0.006 ($p=0.108$)	0.92	0.83	<0.001
Task 2 MOS	0.619 ($p=0.357$)	0.724 ($p=0.001$)	-0.021 ($p<0.001$)	0.014 ($p=0.049$)	0.003 ($p=0.668$)	0.88	0.74	<0.001
Task 2 SIM	1.782 ($p<0.001$)	0.038 ($p=0.617$)	-0.002 ($p=0.254$)	0.026 ($p<0.001$)	0.002 ($p=0.538$)	0.91	0.80	<0.001

- Prediction accuracy of **quality can** be improved by combining multiple objective metrics.
 - By comparing adjusted R-squared values with the individual Pearson correlation coefficients.
- **Task 2 MOS** has lowest adjusted R-squared values
 - Least explainable by the metrics.
 - Predicting cross-lingual quality is harder.

Spoofing performance assessment

Tandem detection cost function (t-DCF)



Actual class	Prior
Target	π_{tar}
Nontarget	π_{non}
Spoof	π_{spoof}
$\Sigma = 1$	

	Actual class	Tandem decision	Unit cost
a.	Target	REJECT (by ASV)	C_{miss}
b.	Nontarget	ACCEPT	C_{fa}
c.	Spoof	ACCEPT	$C_{\text{fa,spoof}}$
d.	Target	REJECT (by CM)	C_{miss}

Systems with highest t-DCF: two patterns

Attacks that do not fool ASV but CM fails to discriminative (=user inconvenience)

Attacks that fool both ASV and CM (=compromised security)

Task 1				
Team ID	ASV EER	CM EER	VC Model	Vocoder
T06	0.00	14.77	StarGAN	WORLD
T08	0.50	37.97	VTLN + Spectral differential	WORLD
T12	0.00	31.46	ADAGAN	AHOcoder
T14	1.00	61.96	One-shot VC	NSF
T28	34.50	32.70	Tacotron	WaveRNN

Task 2				
Team ID	ASV EER	CM EER	VC Model	Vocoder
T08	0.08	46.64	VTLN + Spectral differential	WORLD
T22	30.82	42.97	ASR-TTS (Transformer)	Parallel WaveGAN
T10	45.55	49.81	PPG-VC (LSTM)	WaveNet
T19	44.00	38.35	VQVAE	Parallel WaveGAN
T23	32.82	53.67	CycleVAE	WaveNet

Take-home messages

Correlation with subjective ratings

1. ASV and ASR: high correlation with subjective rating **Useable**
2. MOSNet: better predictions when trained with ASVspoof 2019 data **Potential**
3. Spoofing countermeasure: less correlation **Potential**

Spoofing threat

Both traditional and neural vocoders require attention

Limitations

All the metrics are at **system level**