INTERSPEECH 2018
SEPTEMBER 2-6 | HYDERABAD, INDIA

Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion

Massimiliano Todisco, *EURECOM, France*
Héctor Delgado, *EURECOM, France*
Kong Aik Lee, *NEC Corporation, Japan*
Md Sahidullah, *Inria, France*
Nicholas Evans, *EURECOM, France*
Tomi Kinnunen, *University of Eastern Finland, Finland*
Junichi Yamagishi, *Univ. of Edinburgh, UK & National Institute of Informatics, Japan*

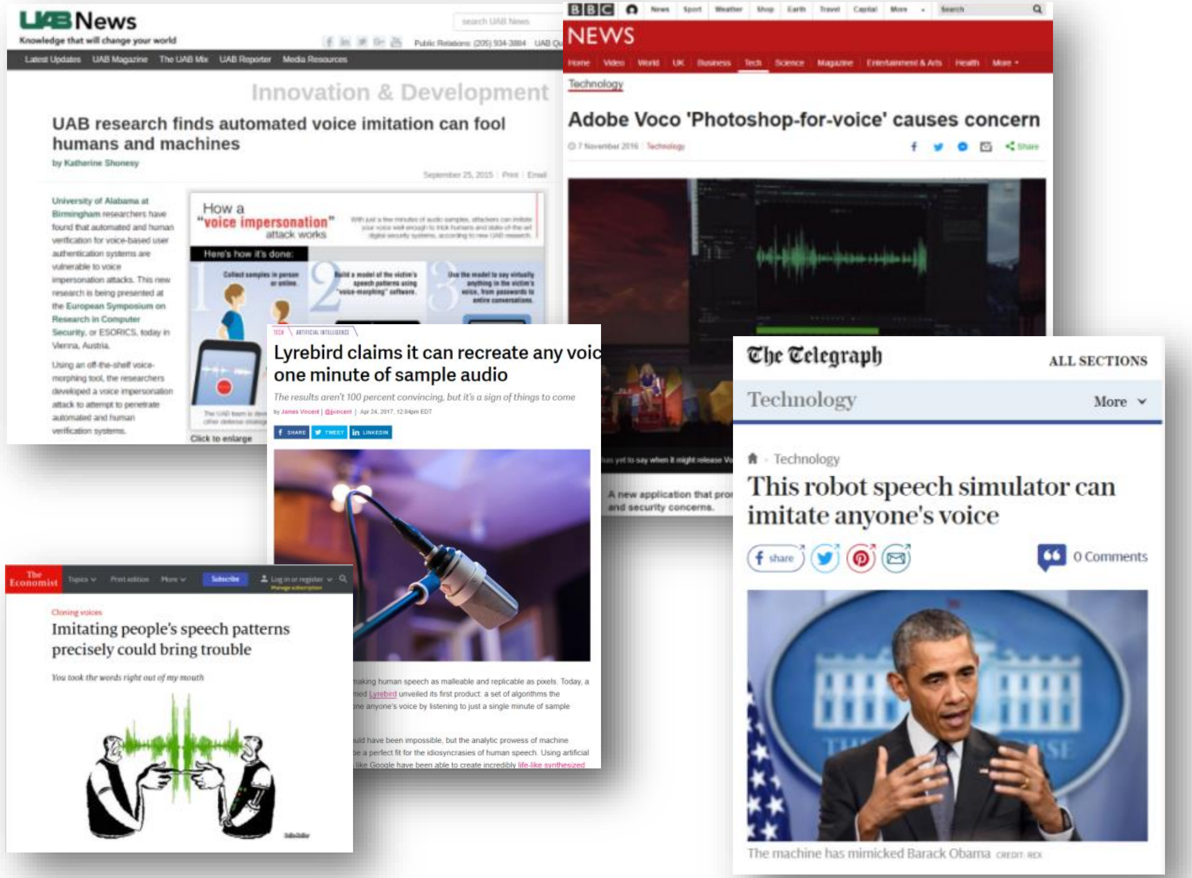
Automatic speaker verification (ASV) and presentation attack (PA)

- Nowadays ASV is a mature approach to person authentication at the level of commercial deployment but...
- ...it is acknowledged that ASV is vulnerable to spoofing (presentation attack (PA) in ISO[1] nomenclature)
- This can undermine confidence in ASV
- Presentation attack detection (PAD) aims **to reduce false acceptances due to spoofing attacks**

[1] ISO standard: ISO/IEC 30107-1:2016

Voice biometric presentation attacks

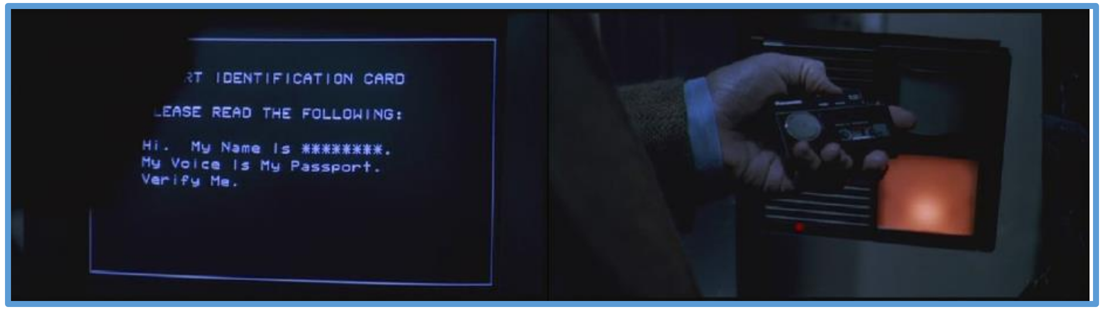
TEXT-TO-SPEECH AND VOICE CONVERSION



REPLAY

Sneakers (1992)

Universal Pictures



Replay device

Device under attack

- *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*

ASVspoof 2015 and 2017 challenge

- ASVspoof initiative provides a standard framework for PAD research
- Standard databases and protocols publicly available
- Focus on logical (2015) and physical (2017) access attacks

- Principal result
  - countermeasures effective in detecting spoofing attacks

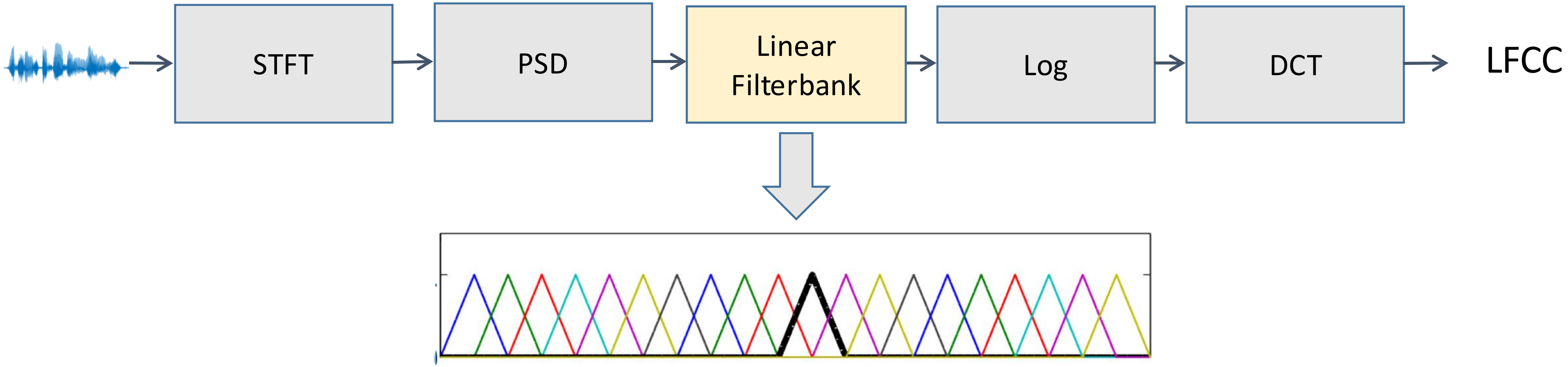
- Missing result:
  - integration of PAD with ASV

Contributions of this work

1. Study and assessment of common front-end for PAD and ASV
2. Gaussian back-end fusion approach to PAD and ASV integration
3. First analysis of this type with ASVspoof 2017 v2.0 database

Common front-end for ASV and PAD: MFCC and LFCC

Linear frequency cepstral coefficients (LFCC)



Common front-end for ASV and PAD: ICMC and CQCC

- Constant Q transform (CQT) [1], [2], is an alternative which employs a variable time/frequency resolution:
  - greater time resolution for higher frequencies
  - greater frequency resolution for lower frequencies

Infinite impulse response Constant Q Mel-frequency cepstral coefficients (ICMC) [3]

→ IIR-CQT → PSD → Mel Filterbank → Log → DCT → ICMC

Constant Q cepstral coefficients (CQCC) [4]

→ CQT → PSD → Log → Uniform resampling → DCT → CQCC

[1] *J. Brown, "Calculation of a constant Q spectral transform," Journal of the Acoustical Society of America, vol. 89, no. 1, pp. 425– 434, January 1991.*

[2] *P. Cancela, M. Rocamora, and E. Lopez, "An efficient multi-resolution spectral transform for music analysis," in Proceedings of ISMIR, 2009, pp. 309–314.*

[3] *H. Delgado, M. Todisco, M. Sahidullah, A. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan; "Further optimizations of constant Q cepstral processing for integrated utterance verification and text-dependent speaker verification"; in proc. of IEEE workshop on Spoken Language Technology, San Diego, CA, December 2016.*

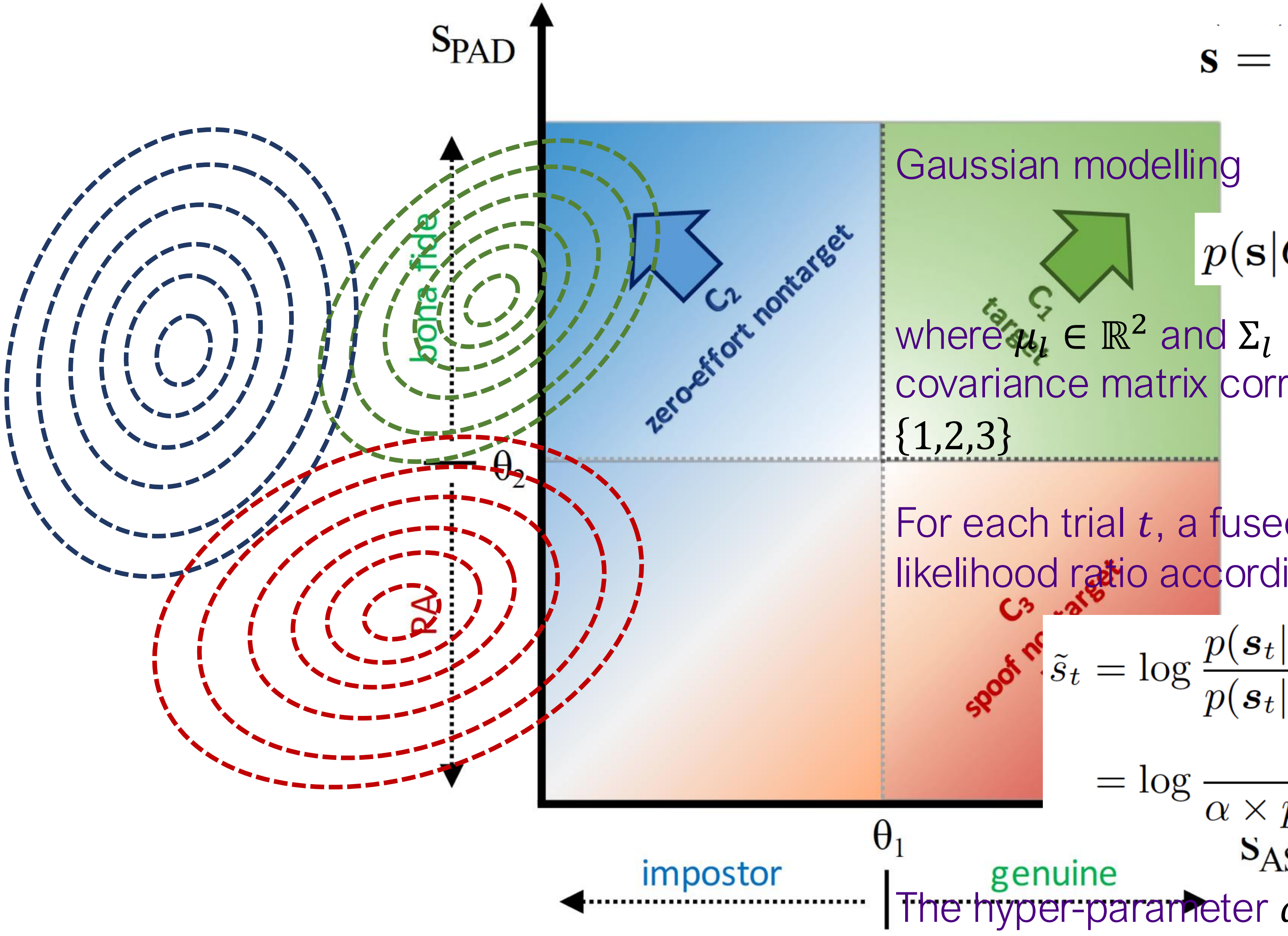
[4] *M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in Speaker Odyssey Workshop, Bilbao, Spain, 2016.*

Integration of PAD with ASV

Definition of positive and negative trials for ASV and CM tasks

class	C_1	C_2	C_3
system/trial	target	zero-effort nontarget	spoof nontarget
PAD	+1	–	-1
ASV	+1	-1	–
ASV + PAD	+1	-1	-1

Integration of PAD with ASV: Gaussian back-end fusion



A two-dimensional view of the score space formed by PAD and ASV

2D feature vector

$$\mathbf{s} = [s_{\mathrm{PAD}}, s_{\mathrm{ASV}}]^{\mathrm{T}} \in \mathbb{R}^2$$

Gaussian modelling

$$p(\mathbf{s}|C_l) = \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$$

where $\mu_l \in \mathbb{R}^2$ and $\Sigma_l \in \mathbb{R}^{2\times 2}$ are the mean vector and the covariance matrix corresponding to class C_l where $l \in \{1,2,3\}$

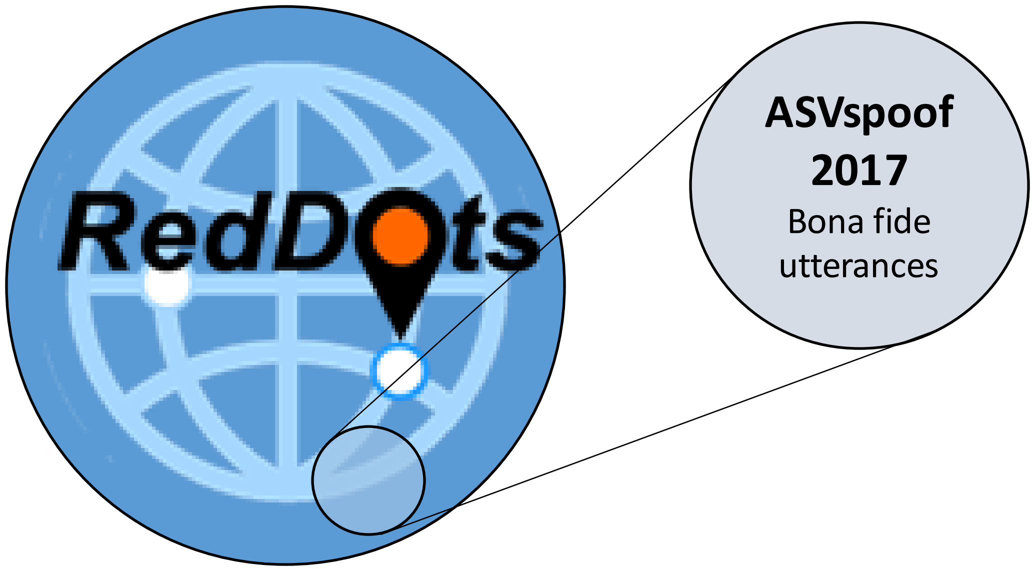
For each trial t, a fused score is computed as the log-likelihood ratio according to

$$\tilde{s}_t = \log \frac{p(\boldsymbol{s}_t|\mathrm{H}_0)}{p(\boldsymbol{s}_t|\mathrm{H}_1)}$$

$$= \log \frac{p(\mathbf{s}_t|C_1)}{\alpha \times p(\mathbf{s}_t|C_2) + (1-\alpha) \times p(\mathbf{s}_t|C_3)}$$

The hyper-parameter $\alpha \in [0,1]$, determines the weight of C_2 and C_3 in the denominator

Experimental setup: ASVspoof 2017 v2.0 ASV/PAD joint protocol



- PAD systems are trained using the official ASVspoof 2017 v2.0 training partition
- Client models for ASV are taken from RedDots part 1 male protocol
- Only trials with matching text are considered
- Joint PAD and ASV experiments were performed using a specifically designed protocol

RedDots corpus
https://sites.google.com/site/thereddotsproject/

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

		#spk	#utterances			#client models
			bona fide		replay	
			target	zero-effort	spoof	
PAD	Training	10	1507		1507	-
PAD & ASV	Dev.	8	742	5186	960	72
	Eval.	24	1106	18624	10878	158

Experimental setup: front-ends, back-ends and ASV/PAD integration

- Standard **MFCC**, **LFCC** and **ICMC** extraction
  - 19 static coefficients (excluding the 0-th)
  - RASTA filtering – (D)elta and (A)cceleration coefficients

- **CQCC** extraction
  - 29 static coefficients (excluding the 0-th)
  - ARTE filtering [1] – (D)elta and (A)cceleration coefficients

- **Log-energy** and **cepstral mean and variance normalization** (CMVN) [2]
- **No speech activity detection** (SAD)

[1] M. Todisco, H. Delgado, and N. Evans, "Articulation rate filtering of CQCC features for automatic speaker verification," in Proc. INTERSPEECH, 2016.

[2] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: metadata analysis and baseline enhancements," in Proc. Odysssey, Les Sables D'Olonne, France, 2018.

Experimental setup: front-ends, back-ends and ASV/PAD integration

- PAD classifier
  - Gaussian mixture models (GMMs) of 512 components
  - Models are learned for bona fide and spoofed speech with an expectation-maximisation (EM) algorithm
  - Classifier scores for a given test utterance are computed as the log-likelihood ratio between the GMMs for bona fide and spoofed speech

- ASV classifier
  - GMM-UBM system
  - 512-component UBM trained with the RSR2015 database
  - Scores are the log-likelihood ratio given the target model and the UBM

- Gaussian back-end fusion
  - Maximum likelihood is used to obtain the means and covariances of all three classes (C1, C2 and C3) on the training set
  - The value of α is set empirically to 0.96 using a grid search on the development set

Experimental result

Speaker verification performance in terms of EER

impostor type	zero-effort		spoof		average		zero-effort		spoof		average	
feat. config / tested on	D	E	D	E	D	E	D	E	D	E	D	E
	Logistic regression fusion [1]						**Polynomial logistic regression fusion [1]**					
MFCC	3.78	2.42	42.72	31.02	23.25	16.72	3.86	2.50	43.81	35.14	23.84	18.82
LFCC	5.72	2.11	46.41	35.71	26.06	18.91	5.47	2.20	37.64	26.99	21.55	14.60
ICMC	2.67	2.16	43.60	33.59	23.14	17.88	2.60	2.08	37.58	29.31	**20.09**	**15.69**
CQCC	6.02	3.52	38.76	33.17	**22.39**	**18.34**	6.02	7.93	42.67	47.96	24.34	27.94
	Cascaded/tandem combination [2]						**Gaussian back-end fusion**					
MFCC	5.19	5.36	23.07	24.65	14.13	15.00	3.99	3.26	21.02	24.35	12.51	13.81
LFCC	7.28	4.96	22.41	21.28	14.84	13.12	5.71	2.90	21.26	17.98	13.48	10.43
ICMC	4.30	4.92	27.08	27.82	15.69	16.37	3.06	3.71	20.90	22.51	11.98	13.11
CQCC	7.31	8.30	15.71	25.26	**11.51**	**16.78**	6.04	4.71	17.34	18.11	**11.69**	**11.41**

[1] I. Chingovska, A. Anjos, and S. Marcel, "Anti-spoofing in action: Joint operation with a verification system," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, 2013.

[2] M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z.-H. Tan, "Integrated spoofing countermeasures and automatic speaker verification: an evaluation on ASVspoof 2015," in INTERSPEECH 2016, San Francisco, USA, 2016.

Experimental result

Two-dimensional PAD-ASV score representation corresponding to LFCC features

Conclusions

- The first comparative assessment of different front-ends for both automatic speaker verification (ASV) and presentation attack detection (PAD)
- The first Gaussian back-end fusion approach to ASV/PAD integration
- Performance is assessed using the ASVspoof 2017 v2.0 database with a suitable ASV/PAD joint protocol
- The best results for independent ASV or PAD tasks do not give the best performance when systems are combined
- The Gaussian back-end approach to integration is shown to generalise well and gives the lowest equal error rate for the independent evaluation set

Useful link

ASVspoof 2017 database Version 2 and ASV/PAD joint protocol available online!

http://www.asvspoof.org

ASVspoof 2019 – tentative schedule

Development data late	November to early December
Evaluation data	early-to-mid February
Score submission +1 week	
Results	end of February
Interspeech deadline	29th March 2019