ASASVIcomtech: The Vicomtech-UGR Speech Deepfake Detection and SASV Systems for the ASVspoof5 Challenge

J. M. Martín[1], E. Rosello[2], A. M. Gomez[2], A. Álvarez[1], **I. López**[2], A. M. Peinado[2]
[1]Vicomtech, [2]Universidad de Granada

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

UNIVERSIDAD DE GRANADA

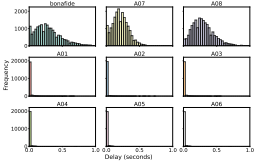
Index

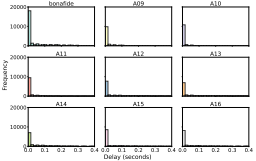
Introduction

- **ASASVIcomtech team**: Researchers from Vicomtech and Universidad de Granada participating in the ASVspoof5 Challenge
- **Track 1**: Speech deepfake detection
  - **Closed condition**: Single system based on a DCCRN architecture
  - **Open condition**: Ensemble of calibrated classifiers based on pre-trained SSL models. Data extension with previous challenges and additional vocoders
- **Track 2**: Spoofing-aware speaker verification
  - **Open condition**: Non-linear late fusion of CM and ASV scores. Best Track 1 system and pre-trained TitaNet ASV model

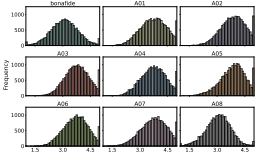
Dataset Analysis

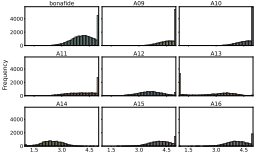
- Analysis of train and dev sets: Balancing, duration, delays, and speech quality
- **Delays**: Amount of time between utterance beginning and speech onset





- **Speech quality**: Non-intrusive ITU-T standard P.563





Track 1 Closed Condition

- **DNN model**: Deep complex convolutional recurrent network (DCCRN)
  - Encoder + LSTM layers (last hidden state) + Softmax layer
  - Weighted cross-entropy loss
- **Input data**: ASVspoof5 dataset
  - 6-second utterances (random segment)
  - 512-sample STFT (128-sample shift)
- **Results**:

Phase	minDCF	actDCF	C_{llr}	EER (%)
Progress	0.4591	1.0000	1.0426	18.63
Evaluation	0.6598	1.0000	1.1159	28.41

Track 1 Open Condition: System

- **DNN models**: Pre-trained SSL models as feature extractors
  - **Upstream**: Wav2Vec2-Large and WavLM-Base (Librispeech corpus)
  - **Adapter**: Weighted sum of different Transformer layers
  - **Downstream**: Per-frame non-linear transform + attentive statistical pooling + cosine scoring
  - **Training**: One-class softmax loss
- **Training data**: Extending ASVspoof5 corpus with external databases
  - **ASVspoof 2019** dataset (train and dev sets, 6 attacks)
  - ASVspoof 2019 **Voc.v4**: 4 additional vocoders

Track 1 Open Condition: System

- **Data augmentation**: On-the-fly during training
  - Trimming leading/trailing silences
  - **RawBoost**: Linear and non-linear convolutive noise, impulsive signal-dependent additive noise, and stationary signal-independent additive noise
- **Calibration**: Output LLR scores. **Beta** calibration trained on the asv5 dev set:

$$L(s') = a \cdot \log \frac{s'}{1 - s'} + c, \quad s' \in [0, 1], \quad a \geq 0, \quad c \in \mathbb{R}$$

- **Ensemble**: Linear weighted sum of W2V2 and WavLM scores

Track 1 Open Condition: Results

- Progress phase

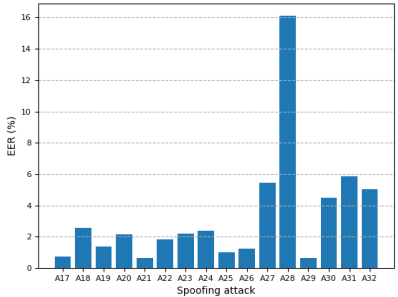
Model	Data	Calibrator	Result			
			minDCF	actDCF	C_{llr}	EER (%)
W2V2	asv19	–	0.3419	0.9818	0.7119	11.79
	asv19voc	–	0.2513	0.9983	0.7220	8.75
	asv5	–	0.0550	0.2679	0.5671	2.02
	asv5	LogReg	0.0550	**0.0563**	0.2000	2.02
	asv5	Beta	0.0550	0.0762	0.1440	2.02
	asv5+19voc	–	0.0354	0.5647	0.5720	1.23
	asv5+19voc	LogReg	0.0354	0.0699	0.1711	1.23
	asv5+19voc	Beta	0.0354	0.0893	0.1254	1.23
WavLM	asv5	–	0.0820	0.2271	0.5597	3.15
	asv5+19voc	–	0.0319	0.0661	0.5048	1.16
	asv5+19voc	Beta	0.0319	0.1423	0.2092	1.16
Ensemble	asv5+19voc	–	**0.0186**	0.2385	0.5368	**0.65**
	asv5+19voc	Beta	**0.0186**	0.0843	**0.1133**	**0.65**

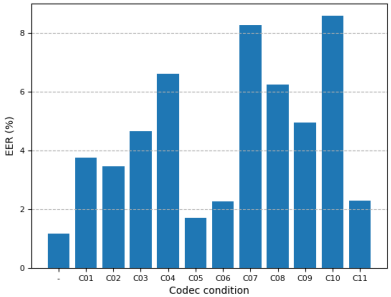
- Evaluation phase

minDCF	actDCF	C_{llr}	EER (%)
0.1348	0.2170	0.3096	5.02

Track 1 Open Condition: Ablation Study

- Analysis in terms of different **spoofing attacks** and **codec conditions**
  - A28: Pre-trained YourTTS | A27, A30-32: Malacopula adversarial attack
  - C04: Encodec | C07: MP3+Encodec | C08: Opus 8 kHz | C10: Speex 8 kHz





Track 2 Open Condition: System

- **ASV subsystem**: Pre-trained TitaNet-Large model (Nvidia NeMo toolkit)
  - **Datasets**: VoxCeleb1&2, Librispeech, telephonic data
  - **Target speaker**: Average speaker embeddings from enrollment utterances
  - **Backends**: Cosine scoring, PLDA (asv5 bonafide train set)
  - **LogReg calibration** (asv5 dev set)
- **CM subsystem**: Calibrated ensemble system from Track 1
- **SASV fusion**: Non-linear fusion LLRs based on LogSumExp function

$$L_{\text{sasv}} = -\log\left[pe^{-L_{\text{asv}}} + (1-p)e^{-L_{\text{cm}}}\right]$$

Track 2 Open Condition: Results

- Progress phase

CM	ASV Backend	Fusion	Calib.	Result		
				min a-DCF	min t-DCF	t-EER (%)
WavLM	Cosine	Linear	×	0.1700	0.1240	4.08
	Cosine	Linear	×	0.1436	0.1102	3.96
	Cosine	LSE ($p = 0.5$)	✓	0.0708	**0.1093**	3.97
Ensemble	Cosine	LSE ($p = 0.7$)	✓	**0.0661**	**0.1093**	3.97
	PLDA	LSE ($p = 0.5$)	✓	0.0752	**0.1093**	**3.83**
	PLDA	LSE ($p = 0.7$)	✓	0.0682	**0.1093**	**3.83**

- Evaluation phase

min a-DCF	min t-DCF	t-EER (%)
0.1295	0.4372	5.43

Conclusions

- Robust ensemble systems based on **pre-trained SSL and ASV** models for Track 1 and 2 open condition
- **Calibration** as a key aspect to compute LLR scores, especially in **SASV fusion**
- **Future work**: Robustness to SOTA synthesis models, novel data augmentation techniques covering new codecs and adversarial attacks
- **Additional details**: Check our Interspeech paper presentation
  - Martín-Doñas *et al.* Exploring Self-Supervised Embeddings and Synthetic Data Augmentation for Robust Audio Deepfake Detection
  - A4-O3: Foundational Models for Deepfake and Spoofed Speech Detection
  - Tuesday 03/09 16:20-16:40

Thank you!

Contact: Juan M. Martín (jmmartin@vicomtech.org)