



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

Juan M. Martín-Doñas¹, Eros Roselló², Angel M. Gomez²,
Aitor Álvarez¹, Iván López-Espejo² and Antonio M. Peinado²

¹Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), San Sebastián, Spain

²Dept. of Signal Theory, Telematics and Communications – CITIC, University of Granada, Spain

{jmmartin, aalvarez}@vicomtech.org, {erosrosello, amgg, iloes, amp}@ugr.es

Abstract

This paper presents the work carried out by the ASASVicomtech team, made up of researchers from Vicomtech and University of Granada, for the ASVspoof5 Challenge. The team has participated in both Track 1 (speech deepfake detection) and Track 2 (spoofing-aware speaker verification). This work started with an analysis of the challenge available data, which was regarded as an essential step to avoid later potential biases of the trained models, and whose main conclusions are presented here. With respect to the proposed approaches, a closed-condition system employing a deep complex convolutional recurrent architecture was developed for Track 1, although, unfortunately, no noteworthy results were achieved. On the other hand, different possibilities of open-condition systems, based on leveraging self-supervised models, augmented training data from previous challenges, and novel vocoders, were explored for both tracks, finally achieving very competitive results with an ensemble system.

1. Introduction

Recent advances in deep learning techniques for the generation of synthetic speech mimicking the voice of a certain speaker poses a challenging threat to our society [1]. These generative models, driven by text-to-speech synthesis (TTS) and voice conversion (VC) technology, have legitimate applications [2] but can be misused for malicious purposes, such as forging voice deepfakes with the aim of blackmailing or vilifying somebody [3]. Moreover, these models can be used in authentication systems based on automatic speaker verification (ASV) to supplant a given user by means of spoofed speech [4].

The scientific community has responded to this situation by a series of challenges which have set up unified development frameworks, thus allowing the establishment of benchmarks and making it possible quick comparisons among different countermeasure (CM) systems. Examples are the ASVspoof challenge series 2015–21 [5], the Audio Deepfake Detection (ADD) challenges 2022–23 [6, 7], and the Spoofing-Aware Speaker Verification (SASV) Challenge 2022 [8].

This year, the fifth ASVspoof challenge (ASVspoof5) was launched and recently wrapped up [9]. This time, the challenge was organized in two tracks: 1) standalone spoofing and speech deepfake detection (non-ASV), and 2) spoofing-aware automatic speaker verification (SASV), where participants could develop their own joint ASV-CM systems. For both tracks, two conditions were considered: closed (developments restricted to ASVspoof5 training data), and open (external data and pre-trained models were also allowed).

This paper presents the work carried out by the ASASVicomtech team, comprised of researchers from Vicomtech and the University of Granada, for the ASVspoof5 Challenge. The team has participated in both Track 1 (closed and open conditions) and Track 2 (open condition only).

For the closed condition of Track 1, we applied a deep complex convolutional recurrent network (DCCRN) fed with full-spectrum features derived from the short-time Fourier transform (STFT). To adapt the DCCRN to this task, we only utilized the CNN encoder and recurrent LSTM layers, omitting the decoder part. Thus, the last LSTM state is projected onto an embedding and then passed through a softmax layer for classification.

For the open condition of Tracks 1 and 2, the team has proposed an ensemble system based on two self-supervised models (Wav2Vec2-Large [10] and WavLM-Base [11]) as deep feature extractors for the CM part. Downstream classifiers are then fine-tuned to compute the CM scores from these deep features. To obtain the ASV scores required for Track 2, we have considered the TitaNet-Large ASV model [12] for embedding extraction and cosine scoring. The final SASV scores are achieved from the calibrated log-likelihood ratio (LLR) ASV and CM scores via non-linear fusion.

The rest of this paper is organized as follows. Section 2 outlines the preliminary analysis performed on the training and validation data. Then, we describe the corresponding systems and challenge results for closed and open conditions in Sections 3 and 4, respectively. Finally, the paper concludes with a summary of the work in Section 5.

2. ASVspoof5 dataset analysis

Before designing our systems for the ASVspoof5 Challenge, we conducted a preliminary analysis of the new training and development datasets. This analysis guided several key decisions in our design process, making it pertinent to include our findings in this paper. In this section, we will examine the database provided for the challenge, focusing on data balance, utterance duration, delays, and speech quality distributions.

2.1. Balancing

Training: The ASVspoof5 training partition contains a total of 182,357 utterances, with 18,797 labeled as bonafide and the remaining samples, as spoofed. Each spoofing attack type (A01-A08) includes 20,445 samples. Therefore, approximately 10% of the training data is bonafide. The dataset is gender-balanced, with roughly 50% of the utterances being male-voiced (92,236 utterances) and 50% female-voiced (90,121 utterances).

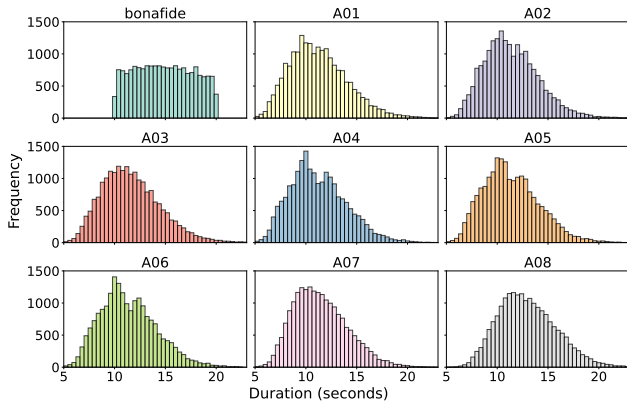


Figure 1: *Histograms of utterance duration from the training set.*

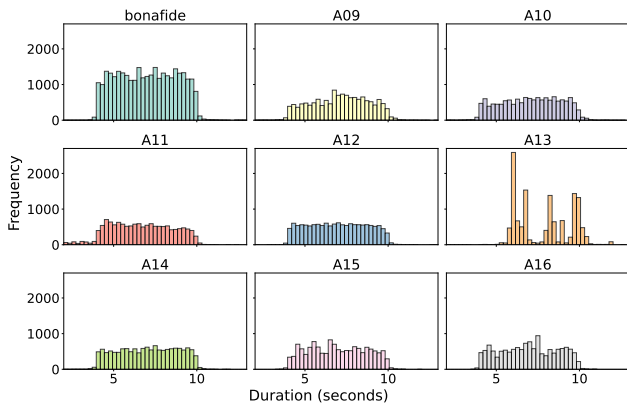


Figure 2: *Histograms of utterance duration from the development set.*

Development: The ASVspoof5 development partition includes 140,950 utterances, with 31,334 being bonafide and the rest, spoofed. Each spoofing attack type (A09-A16) comprises 13,702 samples. In contrast to the training set, approximately 22% of the development data is bonafide. This differs from previous ASVspoof editions, such as that of 2019, where the training and validation datasets had a similar percentage of bonafide audio samples. Like the training set, the development set is also gender-balanced, with about 50% male utterances (71,863 utterances) and 50% female utterances (69,087 utterances).

2.2. Duration of the utterances

Training: The average duration of utterances in the training dataset is 11.92 s, with a standard deviation of 2.99 s. This indicates that the audios contain significantly more information compared to those of the 2019 ASVspoof Challenge, where the training and validation dataset utterances had an average duration of around 4 s. As can be seen from Figure 1, the duration of bonafide audios seems to follow a uniform distribution, ranging from 10 to 20 s. In contrast, the duration of the different attack categories approximates a skewed normal distribution with a mean close to 11 s. This bias in audio length distributions may be leveraged by detection models.

Development: In the development dataset, however, the audios are shorter, with an average duration of 7.08 s and a standard

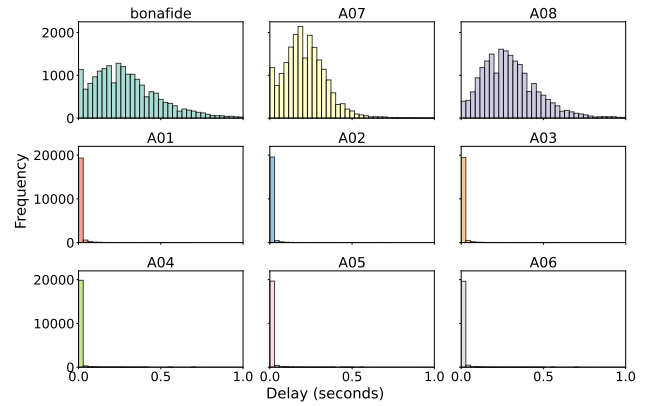


Figure 3: *Histograms of utterance delay from the training set.*

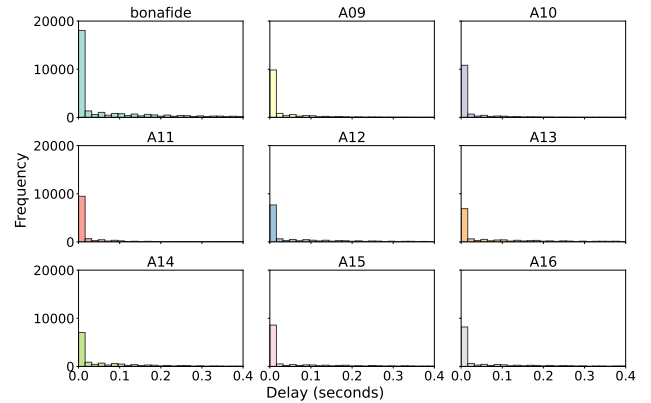


Figure 4: *Histograms of utterance delay from the development set.*

deviation of 1.84 s. As depicted in Figure 2, the distributions also differ, with most classes following a distribution similar to a uniform one ranging from 3 to 10 s. Exceptions include attack A13, which lacks a clear distribution pattern, and attack A11, which contains some audios with a duration less than 3 s.

2.3. Delays

Since initial silences were found to introduce a bias in the ASVspoof 2019 database [13], we also analyzed the utterance delay¹ distributions in the ASVspoof5 datasets. We employed a voice activity detector to spot the onset of speech and, subsequently, calculate the delay on an utterance basis.

Training: In the training dataset, the vast majority of audios appear to have no delay. However, it is important to note that those with delays are either bonafide samples or attack types A07 and A08, as Figure 3 shows. This bias could potentially be exploited by a model to classify the audios, leading to overfit the training dataset.

Development: In the development set, delays are almost negligible, with a short duration, and do not appear to be associated with any particular class. This contrasts with the case of the training set, where trimming or similar approaches could be advised.

¹In this work, delay is defined as the amount of time between the beginning of an audio file and the speech onset.

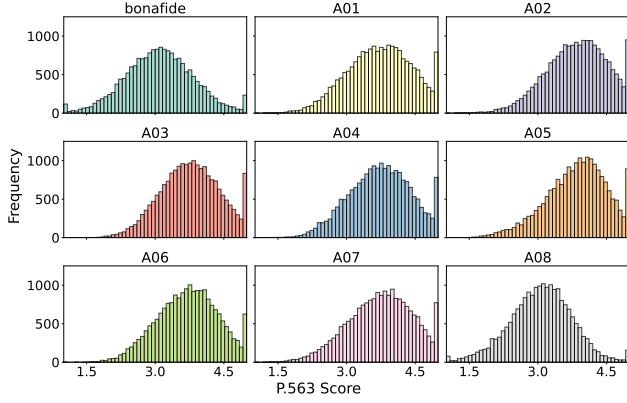


Figure 5: Histograms of P.563 scores for training utterances across different attack types.

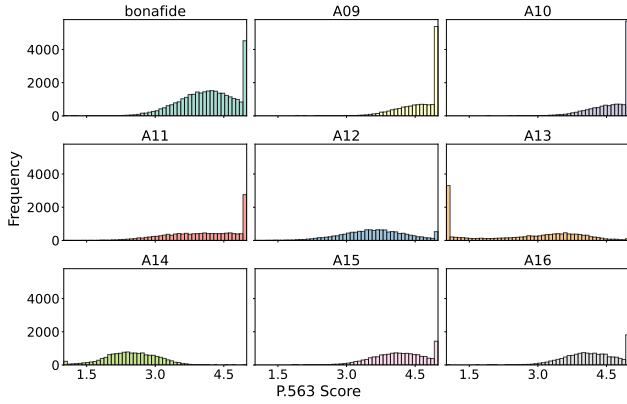


Figure 6: Histograms of P.563 scores for development utterances across different attack types.

2.4. Speech quality

Finally, in addition to the previous analyses, we conducted a speech quality evaluation on the training and validation datasets by means of the non-intrusive ITU-T standard P.563 [14] as an objective perceptual quality metric.

Training: On the training data, the speech quality scores seem to follow a quasi-normal distribution, with some clustering at the maximum value. The bonafide class has a slightly lower mean compared to most attack classes, as shown in Figure 5. Overall, the data show a diverse range of speech quality scores.

Development: In contrast to the above, the validation set appears to be less varied, exhibiting a more limited range and a tendency for values to cluster at the maximum. As illustrated by Figure 6, there are two attacks, A13 and A14, with significantly lower quality metric values than the others. This pronounced difference may make these attacks easier to be distinguished from the other ones.

The disparity between the whole training and validation datasets is clearly evident from Figure 7. Although there is some clustering at the maximum value for the training data, the overall distribution is more varied compared to that of the validation dataset. Furthermore, the validation dataset tends to accumulate more heavily at the maximum and minimum values, indicating less diversity in speech quality.

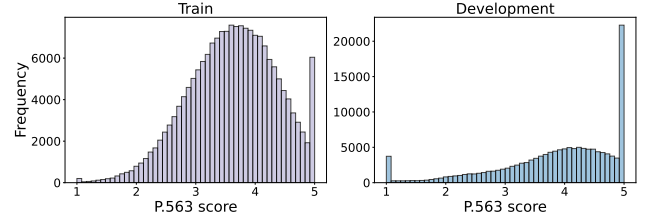


Figure 7: Training and development utterance P.563 score histograms.

3. Closed-condition system

Our team participated in the closed condition with a novel single system which, unfortunately, did not reach our expectations. Due to this, the system was only evaluated on Track 1. For completeness sake, the system is briefly described below.

3.1. System description

DNN model: Our proposal is derived from the deep complex convolutional recurrent network (DCCRN) first described in [15] for speech enhancement tasks. This network consists of a causal convolutional encoder-decoder architecture with LSTM layers between the encoder and the decoder, so that the temporal dependencies can be modeled. We chose this network because of its ability to process full spectra (i.e., both magnitude and phase), as the DCCRN is essentially an extension of a CRN that performs joint complex-value computation instead of considering two isolated real and imaginary parts. For deepfake detection, we have removed the decoder part of the architecture. Thus, the last hidden state of the deepest LSTM layer is used to compute an embedding from the received input, which is finally mapped into classes by a softmax layer.

Input data: We set a fixed input of 96,000 samples (6 s), so that the network does not exploit any length bias (see Section 2). Similarly, audio excerpts are extracted from the middle of the utterance during validation and testing, and from a random position during training, in order to prevent the network from learning of the initial delays. A complex-spectrum representation is then obtained via STFT [16] with a square-root Hann window of 512 samples shifted by 128 (32 and 8 ms at 16 kHz). Thus, input tensors are of size $(C \times F \times T)$, with $C = 2$ channels (real and imaginary parts), $F = 256$ bins of frequency (DC component is removed) and $T = 750$ frames.

Training setup: Network optimization is performed through the ADAM algorithm [17] with a batch size of 64 utterances and a learning rate of $3 \cdot 10^{-4}$. We use the weighted cross entropy (WCE) [18] as a loss function during training. Despite the excellent convergence on the training dataset, WCE scores were very pathological during validation, diverging after the first epoch. As this issue could not be resolved by adjusting the learning rate or changing the training loss function, we trained the model for 100 epochs and selected the one that achieved the best EER on the validation dataset (provided it was not a transient spike).

3.2. Challenge results

As mentioned above, model performance was not satisfactory, mainly due to the validation issues found and the apparent inability of the network to generalize. Table 1 summarizes the results obtained during the progress and evaluation phases of

Table 1: *Track 1 closed-condition results provided by the DC-CRN model.*

Phase	minDCF	actDCF	C_{llr} [20]	EER (%)
<i>Progress</i>	0.4591	1.0000	1.0426	18.63
<i>Evaluation</i>	0.6598	1.0000	1.1159	28.41

the challenge. As can be observed, our proposal improves the AASIST [19] baseline, provided by the organizers [9], in terms of minDCF (0.6598 vs. 0.7110), but EER is marginally the same (28.41% vs. 29.12%).

4. Open-condition systems

4.1. System description

Our team also participated in the open condition with an ensemble system that yielded significantly better results. This system demonstrated improved performance and robustness in both Track 1 and 2. Below, we provide a description of this successful system.

4.1.1. Track 1: Speech deepfake detection

DNN models: Our approach is based on pre-trained self-supervised learning (SSL) speech models as feature extractors to obtain robust deep embeddings. We considered two different SSL models pre-trained on the LibriSpeech corpus: Wav2Vec2-Large (W2V2) [10] and WavLM-Base (WavLM) [11]. The SSL deep embeddings are then converted to a final spoof score through a downstream model fine-tuned on the target data. This downstream model is different depending on the corresponding SSL upstream considered. This is because padding masks are used in WavLM (and surrogate downstream) during training, but not in the W2V2 model, following model recommendations. Therefore, the classifier for our winner system in ADD 2022 Track 1 [21] is combined with the W2V2 feature extractor. On the other hand, for WavLM upstream, we chose the NN-ASP classifier from our recent paper [22], which takes into account the padding mask. Nevertheless, the structure of those downstreams is similar, following weighted sum of the Transformer layers, per-frame non-linear transformations, attentive statistical pooling, and cosine scoring. These downstreams are trained by minimizing the one-class softmax loss [23].

Train and development data: The main corpus for training is the ASVspoof5 dataset previously described. To boost deepfake detection and robustness against different attacks, we extended this corpus with external databases. Thus, we included the training and development data from the 2019 ASVspoof Challenge [24], based on the VCTK database. Moreover, we extended this dataset by aggregating additional vocoded data as described in [25]. We used the `Voc.v4` partition, which considers four additional vocoders pre-trained on LibriSpeech and fine-tuned on ASVspoof 2019 bonafide speech. The train and development sets from the in- and out-domain data are not mixed, i.e., original train data are only augmented with the train sets of the out-domain datasets (similar with development partitions).

Data augmentation: We applied on-the-fly data augmentation techniques during the training of our systems. First, we trimmed the leading and trailing silences of the signals to avoid exploiting potential misleading artifacts from the databases, especially from the ASVspoof 2019 dataset. Then, to emulate the effect of different codec systems on the clean speech sig-

nals, we applied RawBoost data augmentation [26]. In this case, we considered the full configuration of RawBoost including three kinds of distortions: linear and non-linear convolutive noise, impulsive signal-dependent additive noise, and stationary signal-independent additive noise.

Training setup: The models are fine-tuned on the corresponding training data using the ADAM optimizer [17] with a learning rate of $3 \cdot 10^{-4}$. The effective batch size is 64 utterances (8 utterances mini-batch and 8 steps for gradient accumulation). On the other hand, the development data are only considered for best model selection and early-stopping.

Calibration: The output cosine scores of the detection systems are not well-calibrated LLRs, making those scores sub-optimal for proper Bayesian decisions [20]. To reduce the calibration loss, we trained an additional calibration backend using the ASVspoof5 development set. The calibrator is only restricted to be a monotonic rising function converting the raw scores to LLRs calibrated on the development set. Apart from the well-known logistic regression (LogReg), defined as a linear function for LLR scores, we considered the beta calibration [27], which assumes beta distributions for the scores. Beta calibration seems better suited for cosine scores within a specific range. We applied the univariate version of the calibration function, defined as $L(s') = a \cdot \log \frac{s'}{1-s'} + c$, where s' are the scaled scores in the range $[0, 1]$, while $a \geq 0$ and $c \in \mathbb{R}$ are the function parameters to be fitted. It should be noticed that this calibrator can be trained as a LogReg by first converting the scaled scores through the logarithmic function.

Ensemble system: Finally, we also explored the ensemble of our best-performing detection systems through late score fusion. To this end, the output scores from both the W2V2 and WavLM subsystems are combined by a linear weighted sum. We evaluated the fusion of both raw scores and calibrated LLRs.

4.1.2. Track 2: Spoofing-aware ASV

For the SASV open track, we proposed a straightforward combination of ASV and CM subsystems by score fusion. This allowed us to quickly deploy and evaluate an SASV system based on a fixed ASV model and our best CM system for Track 1.

For the ASV subsystem, we considered the pre-trained TitaNet-Large model included in the Nvidia NeMo toolkit [12]. This architecture is based on a convolutional network with squeeze-and-excitation layers and channel attention to extract a speaker embedding from an utterance. This network is trained with multiple ASV corpora, including VoxCeleb 1 and 2, LibriSpeech, and telephonic data (NIST SRE 04-08, Fisher and Switchboard). The embedding for the target speaker in each trial is obtained by averaging the embeddings from the different enrollment utterances of the speaker. Two different scoring backends are evaluated: cosine scoring, and a probabilistic linear discriminant analysis (PLDA) model trained on the ASVspoof5 bonafide training data. Moreover, the output scores are also calibrated on the ASVspoof5 development set (considering target and non-target trials only) by means of LogReg. This calibration is especially important for proper integration with the calibrated CM scores.

Finally, for score fusion, we compared two different approaches. The first one is simply a linear weighted sum of the scores. A better procedure is proposed in [28] based on a non-linear fusion of LLR scores. To this end, a negative LogSumExp (LSE) function (smooth maximum) is applied to the negative LLR scores, yielding the final SASV LLR scores. Furthermore, a different weight is considered for the scores during the sum,

Table 2: Track 1 open-condition results for the different systems proposed during the 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

Table 3: Evaluation phase results for our submitted system in Track 1 open condition.

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

which is adjusted by a grid search on the development data. Note that, in contrast to [28], our approach does not fit Gaussian distributions to the raw score vectors. Instead, we directly use the LLRs from the individual subsystems, thereby avoiding potential overfitting to the training/development datasets.

4.2. Challenge results

4.2.1. Track 1

We first evaluated the different configurations for our proposed approach on the progress subset. Table 2 shows the results achieved for different combinations of DNN models, training data and calibration method. With respect to the data, we can observe that the combination of ASVspoof5 and ASVspoof 2019 generally yields the best results, and using additional vocoding data is beneficial for generalization purposes (comparing asv19 and asv19voc experiments). Calibration also helps improve secondary metrics, with beta calibration giving better results through different operation points (lower C_{llr} values). WavLM models produce better results than W2V2 when the aggregated data are used for training, and the model ensemble fusion yields the best performance. Thus, we considered this ensemble model with beta calibration as our final system for the subsequent evaluation phase.

Our final results for the evaluation phase are shown in Table 3. We achieve competitive results with a 0.1348 minDCF and an EER of 5.02% while also keeping good performance in terms of the other two metrics that measure both discrimination and calibration capabilities. To disentangle these results, we show EER values broken down by spoofing attack and codec in Table 4. We choose EER instead of the primary metric minDCF since EER makes the comparison easier. That being said, both metrics are directly related, and similar trends can be observed. Results in terms of the attack type reveal that our system is

mainly negatively affected by A28 (16.10% EER). This spoof attack corresponds to a pre-trained YourTTS model [32]. It is interesting to note that a similar attack using YourTTS is also included in the development set, where this performance degradation was not observed, which can be due to different configurations or modifications. Further investigations will be needed to comprehend this difference. We can also see that attacks A27 and A30-32 yield a higher EER (~5%) in comparison with other spoof attacks. The common feature of those spoofing systems is that they also include the Malacopula adversarial attack [33]. Although our approach generally behaves properly against adversarial attacks, further countermeasures should be taken into account to improve the results under A27 and A30-32. On the other hand, results across codec conditions show that the good performance in clean conditions (1.17% EER) is severely degraded when using codecs C07 (MP3+Encodect) and C10 (Speex 8 kHz), and moderately degraded by C04 (Encodect) and C08 (Opus 8 kHz). In general, our approach is more affected under narrowband conditions (8 kHz) and neural audio compression (Encodect [34]). Although the RawBoost data augmentation helps better generalize across codec conditions, it cannot completely cover these two scenarios, probably requiring additional augmentation techniques that can cope with the new degradations and artifacts produced by these channel codecs. Nonetheless, it can be observed that the performance of our approach is generally robust and competitive across a broad set of codecs and spoofing attacks.

4.2.2. Track 2

Table 5 depicts our results for the Track 2 open condition during the progress phase. Due to the limited amount of trials, we mainly considered our best fusion configurations evaluated on the development set, which are based on LSE score fusion from calibrated systems. This table also includes linear fusion using cosine scores with both WavLM and ensemble CM as baselines. As can be observed, using LSE fusion with calibrated LLRs yields better performance than a simple linear fusion, especially when considering the minimum a-DCF metric (a-DCF from now on). Adjusting the weight between CM and ASV scores can also improve the results. In our case, higher weights p for the ASV scores produced, in general, better performance. Finally, a comparison between ASV backends reveals that sim-

Table 4: Track 1 EER (%) results achieved by our open-condition system, broken down by spoofing attack and codec condition.

	–	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	Pooled
A17	0.05	0.20	0.17	0.37	0.50	0.05	0.15	1.00	1.30	0.43	0.79	0.12	0.72
A18	0.15	1.47	1.38	1.81	3.35	0.37	0.68	4.43	2.85	2.70	5.13	0.38	2.58
A19	0.10	0.66	0.51	0.91	2.13	0.14	0.27	2.74	1.35	0.49	1.30	0.14	1.38
A20	0.15	1.06	1.05	1.76	3.36	0.24	0.64	5.25	2.77	1.49	4.32	0.18	2.16
A21	0.01	0.29	0.17	0.41	0.92	0.04	0.09	1.22	1.13	0.24	0.71	0.05	0.64
A22	0.10	0.84	0.67	1.65	3.15	0.25	0.44	4.55	2.18	1.55	2.92	0.21	1.84
A23	0.13	1.01	1.16	1.94	3.48	0.29	0.59	5.02	2.34	1.79	3.79	0.21	2.22
A24	0.21	1.14	1.05	2.48	3.61	0.33	0.58	5.15	2.59	1.31	3.63	0.29	2.39
A25	0.05	0.30	0.33	0.62	1.66	0.12	0.15	2.54	1.59	0.49	1.54	0.05	1.02
A26	0.05	0.42	0.34	1.11	2.12	0.09	0.19	2.67	1.76	0.68	1.96	0.12	1.26
A27	0.41	3.36	3.97	6.38	8.13	0.84	1.90	11.02	7.32	6.22	12.46	0.63	5.45
A28	4.99	13.57	12.09	13.90	20.34	6.57	9.18	24.21	22.15	18.15	26.66	10.60	16.10
A29	0.15	0.50	0.42	0.58	0.54	0.33	0.34	0.58	1.00	0.53	0.50	0.42	0.63
A30	0.37	2.99	2.70	3.91	6.07	0.75	1.42	8.18	5.52	4.44	9.16	0.50	4.47
A31	0.79	4.40	4.61	6.63	10.21	1.27	1.91	12.86	6.97	6.32	10.73	0.75	5.87
A32	0.29	2.94	3.23	5.47	7.67	0.68	1.64	9.70	7.22	5.78	12.46	0.69	5.02
Pooled	1.17	3.75	3.46	4.67	6.60	1.70	2.27	8.28	6.24	4.95	8.58	2.29	5.02

Table 5: Track 2 open-condition results for the different systems evaluated during the progress phase.

CM	ASV Backend	Fusion	Calib.	Result		
				min a-DCF [29]	min t-DCF [30]	t-EER (%) [31]
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

Table 6: Evaluation phase results for our submitted system in Track 2 open condition.

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

ilar performances are achieved, with the cosine (PLDA) scoring outperforming in terms of a-DCF (t-EER). This demonstrates that the pre-trained ASV embedding extractor is competitive enough to obtain good verification performance. Thus, we selected cosine scoring as our ASV backend, and combined the calibrated LLR scores from CM and ASV subsystems using the LSE score fusion.

Finally, we present our results for the evaluation phase in Table 6. We achieved a competitive a-DCF of 0.1295 with our proposed system, as well as a strong performance in terms of other tandem-related metrics. Regarding results per attack type and codec condition, we observed similar trends to those from our Track 1 results. The most challenging attack was A28 (0.451 a-DCF), followed by the systems using the Malacopula adversarial attack (results within the range [0.111, 0.153]). Moreover, the most challenging codecs were C08 and C10, with results close to 0.195 a-DCF, especially compared to clean conditions (0.055 a-DCF). Nevertheless, we achieved a robust and straightforward score fusion approach based on reliable ASV and CM subsystems, resulting in competitive challenge results.

5. Conclusion

In this paper, we have presented the Vicomtech-UGR systems submitted to the ASVspoof5 Challenge. After facing difficulties in developing CM systems for the Track 1 closed condition, we achieved a robust ensemble system with competitive performance in the open condition. This was due to leveraging self-supervised models, and augmented training data from previous challenges and novel vocoders. For the SASV system of Track 2, we have combined our ensemble CM system with a pre-trained ASV model via a straightforward non-linear score fusion. For both tracks, calibration has been a key aspect to provide meaningful LLR scores, especially during the integration of ASV and CM subsystems. As future work, we will analyze the robustness of our speech deepfake detection approach against state-of-the-art speech synthesis models, and the development of advanced data augmentation techniques covering additional codecs, narrowband channels, and adversarial attacks.

6. Acknowledgements

This work is part of the project PID2022-138711OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, and the FPI grant PRE2022-000363. Also, this work has been supported in part by the European Union’s Horizon Europe research and innovation programme in the context of project EITHOS under Grant Agreement No. 101073928.

7. References

- [1] Momina Masood, Marriam Nawaz, Khalid Malik, Ali Javed, Aun Irtaza, and Hafiz Malik, “Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward,” *Applied Intelligence*, vol. 53, pp. 1–53, 2022.
- [2] Jose A. Gonzalez-Lopez, Alejandro Gomez-Alanis, Juan M. Martín-Doñas, José L. Pérez-Córdoba, and Angel M. Gomez, “Silent speech interfaces for speech restoration: A review,” *IEEE Access*, vol. 8, pp. 177995–178021, 2020.
- [3] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado, “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection,” in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.
- [4] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, “Spoofing and countermeasures for speaker verification: A survey,” *Speech Communication*, vol. 66, 2014.
- [5] “ASVspoof: Automatic Speaker Verification and Spoofing Countermeasures Challenge,” 2015, Accessed on July 26th, 2024.
- [6] Jiangyan Yi et al., “ADD 2022: The first audio deep synthesis detection challenge,” in *Proc. ICASSP 2022*, 2022, pp. 9216–9220.
- [7] Jiangyan Yi et al., “ADD 2023: The second audio deepfake detection challenge,” in *Proc. IJCAI 2023 DADA Workshop*, 2023, pp. 125–130.
- [8] Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen, “SASV 2022: The First Spoofing-Aware Speaker Verification Challenge,” in *Proc. Interspeech 2022*, 2022, pp. 2893–2897.
- [9] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi, “ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale,” in *ASVspoof Workshop 2024 (accepted)*, 2024.
- [10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.
- [11] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” *IEEE J-STSP*, vol. 16, pp. 1505–1518, 2022.
- [12] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg, “TitaNet: Neural Model for speaker representation with 1D Depth-wise separable convolutions and global context,” in *Proc. ICASSP 2022*, 2022, pp. 8102–8106.
- [13] Nicolas Müller, Franziska Dieckmann, Pavel Czempein, Roman Canals, Konstantin Böttinger, and Jennifer Williams, “Speech is silver, silence is golden: What do ASVspoof-trained models really learn?,” in *Proc. 2021 ASVspoof Challenge Workshop*, 2021, pp. 55–60.
- [14] International Telecommunication Union, “Single-ended method for objective speech quality assessment in narrow-band telephony applications,” Standard P.563, ITU-T, 2004.
- [15] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” *arXiv:2008.00264*, 2020.
- [16] Leigh D. Alsteris and Kuldip K. Paliwal, “Short-time phase spectrum in speech processing: A review and some experimental results,” *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.
- [17] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *Proc. of 3rd International Conference on Learning Representations*, 2015, pp. 1–13.
- [18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [19] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans, “AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.
- [20] David A. van Leeuwen and Niko Brümmer, “An introduction to application-independent evaluation of speaker recognition systems,” in *Speaker Classification I: Fundamentals, Features, and Methods*, pp. 330–353. Springer Berlin Heidelberg, 2007.
- [21] Juan M. Martín-Doñas and Aitor Álvarez, “The Vi-comtech Audio Deepfake Detection System based on Wav2Vec2 for the 2022 ADD Challenge,” in *Proc. ICASSP 2022*, 2022, pp. 9241–9245.
- [22] Juan M. Martín-Doñas, Aitor Álvarez, Eros Rosello, Angel M. Gomez, and Antonio M. Peinado, “Exploring self-supervised embeddings and synthetic data augmentation for robust audio deepfake detection,” in *Proc. Interspeech 2024 (accepted)*, 2024.
- [23] You Zhang, Fei Jiang, and Zhiyao Duan, “One-class learning towards synthetic voice spoofing detection,” *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.
- [24] Andreas Nautsch et al., “ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.
- [25] Xin Wang and Junichi Yamagishi, “Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders,” in *Proc. ICASSP 2023*, 2023.
- [26] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans, “RawBoost: A Raw Data Boosting and Augmentation Method applied to Automatic Speaker Verification Anti-Spoofing,” in *Proc. ICASSP 2022*, 2022, pp. 6382–6386.

- [27] Meelis Kull, Telmo Silva Filho, and Peter Flach, “Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers,” in *Artificial Intelligence and Statistics*, 2017, pp. 623–631.
- [28] Xin Wang, Tomi Kinnunen, Lee Kong Aik, Paul-Gauthier Noe, and Junichi Yamagishi, “Revisiting and improving scoring fusion for spoofing-aware speaker verification using compositional data analysis,” in *Proc. Interspeech 2024 (accepted)*, 2024.
- [29] Hye jin Shim, Jee weon Jung, Tomi Kinnunen, Nicholas Evans, Jean-François Bonastre, and Itshak Lapidot, “a-DCF: An architecture agnostic metric with application to spoofing-robust speaker verification,” in *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, 2024, pp. 158–164.
- [30] Tomi Kinnunen et al., “Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.
- [31] Tomi H. Kinnunen, Kong Aik Lee, Hemlata Tak, Nicholas Evans, and Andreas Nautsch, “t-EER: Parameter-free tandem evaluation of countermeasures and biometric comparators,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2622–2637, 2024.
- [32] Edresson Casanova, Julian Weber, Christopher D. Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A. Ponti, “YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone,” in *Proc. ICML*, 2022, pp. 2709–2720.
- [33] Massimiliano Todisco, Michele Panariello, Xin Wang, Hector Delgado, Kong-Aik Lee, and Nicholas Evans, “Malacopula: Adversarial automatic speaker verification attacks using a neural-based generalised Hammerstein model,” in *ASVspoof 2024 Workshop (submitted)*, 2024.
- [34] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, “High fidelity neural audio compression,” *Transactions on Machine Learning Research*, 2023.