Dual-Stream Temporal Convolutional Neural Network for Voice Presentation Attack Detection

Lazaro J. Gonzalez-Soler[1], Marta Gomez-Barrero[2], Madhu Kamble[3], Massimiliano Todisco[3], Christoph Busch[1]
1 - Biometrics and Internet Security Research Group
Hochschule Darmstadt, Germany
{lazaro-janier.gonzalez-soler,christoph.busch}@h-da.de
2 - Hochschule Ansbach, Germany
marta.gomez-barrero@hs-ansbach.de
3 - EURECOM, France
{kamble,todisco}@eurecom.fr

***Abstract*—Improving the robustness of biometric systems to external attacks is of the utmost importance for the research community. In particular, Automatic Speaker Verification (ASV) can be easily bypassed by launching either attack presentations (i.e., physical access attacks) over the capture devices (i.e., microphone) or exchanging the input sample in the channel between the capture device and the signal processor (i.e., logical access attacks). In order to address these security threats, ASVspoof challenges have evaluated the generalisation ability of several Presentation Attack Detection (PAD) approaches in the last decade. Those algorithms have reported a remarkable detection performance to detect physical and logical access attacks when they are combined with the decision provided by the ASV systems. They fundamentally depend upon the complementary information of ASV systems for a reliable detection performance. Therefore, they are not interoperable across different systems. In this work, we propose an interoperable dual-stream PAD method which leverages temporal information from image-based voice spectrograms to enhance generalisation on PAD. The experimental results conducted over the publicly available ASVspoof 2019 and 2021 databases show the feasibility of our approach to detect both physical and logical access attacks unknown in training.**

***Index Terms*—Presentation Attack Detection, Automatic Speaker Verification, Generalisation Capability, Temporal Convolutional Neural Network**

I. Introduction

Biometric systems have considerably evolved in recent years, mainly due to the breakthroughs achieved by Deep Neural Networks (DNN). As a result, they have been widely deployed in a variety of applications, such as bank accounts, mobile phone unlocking, and call centres. In spite of their advantages, these systems are still vulnerable to attack presentations (APs) which can be easily launched over the biometric captured device by a non-authorised subject to gain access to the aforementioned applications. This non-authorised individual could also manipulate the biometric characteristics to avoid detection.

This research work has been funded by the DFG-ANR RESPECT Project (406880674), and the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

In order to address these security threats, ASVspoof challenges have significantly promoted the study and analysis of several Presentation Attack Detection (PAD) approaches to protect Automatic Speaker Verification (ASV) in the last decade. The recent ASVspoof 2021 [1] showed that most kinds of Presentation Attack Instruments (PAIs) can be successfully detected by current PAD approaches. However, these techniques fundamentally depend upon the complementary information provided by the ASV systems in the ensemble for a reliable detection performance. Therefore, they are not interoperable across different ASV schemes.

To overcome these limitations, some studies have addressed voice PAD through image representation of spectrograms. Alegre *et al.* [2] proposed a PAD method based on the combination of Local Binary Patterns (LBP) and one-class classifiers. Following this idea, deep residual learning [3], which has been successfully used in image processing tasks, was adopted for voice PAD. The ResNet model is employed in combination with image-like speech spectrograms (e.g., Mel spectrogram) for the detection of voice APs [4], [5]. Recently, Gonzalez-Soler *et al.* [6] explored the feasibility of transforming 1D spectrograms to images through four different techniques and analysed several well-known textural features in combination with the Fisher Vector representation [7] to improve the detection of unknown PAI species. In spite of the efforts carried out, those methods still lack generalisation beyond PAI species on which they are trained.

Motivated by that fact, we propose in this work a dual-stream temporal convolutional network which leverages temporal information of latent embeddings extracted from the two best spectrogram-to-image representations as reported in [6] (i.e., constant Q transform - CQT [8] and short-time Fourier transform - STFT [9]) to enhance the generalisation capabilities. This framework transforms the input raw audio waveform to a 2D image which can feed to a Convolutional Neural Network (CNN). In our experiments, embeddings are obtained from three different traditional CNNs (i.e., DenseNet [10], ResNet [11], and MobileNetv2 (i.e.,

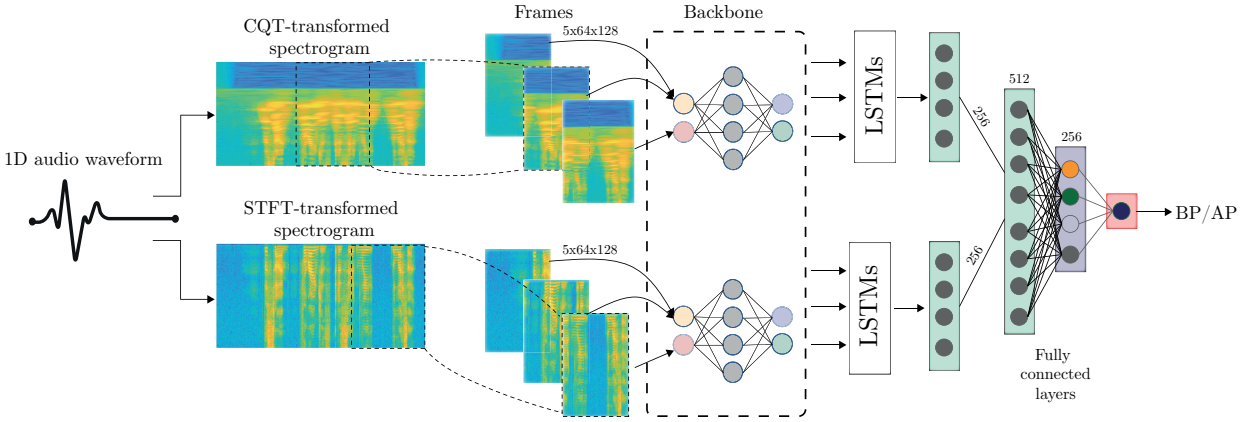


Fig. 1: General overview of our dual-stream temporal CNN approach.

version 2) [12]). The experimental evaluation conducted over the well-known ASVspoof 2019 [13] and 2021 [5] databases in compliance with the international metrics ISO/IEC 30107-3 [14] for biometric PAD shows the soundness of our proposed method to spot PAIs over challenging scenarios.

The remainder of this paper is organised as follows: Sect. II describes our dual-stream temporal CNN for voice PAD. The experimental protocol is explained in Sect. III. The results benchmarking the performance of our proposed method with the state-of-the-art systems are discussed in Sect IV. Conclusions and future work directions are finally summarised in Sect. V.

II. Dual-stream Temporal Convolutional Neural Network

Fig. 1 shows a general overview of our dual-stream temporal CNN which takes advantage of the temporal latent representation of the input spectrograms for voice PAD. In essence, the input 1D audio waveforms are firstly transformed to 2D images using two different strategies (i.e., CQT [8] and STFT [9]). This transformation will lead to an individual stream in our approach. The new images are then split into several frames and represented by an intermediate latent vector stemming from a traditional CNN (e.g., DenseNet, declared as Backbone in Fig. 1). In order to include voice temporal information in the network optimisation, the per-frame latent representation is further processed by a series of Long Short-Term Memory (LSTM) layers, whose output layer is concatenated with the one provided by the other stream. Finally, these final concatenated features are fed to a fully connected (FC) layer, which, in turn, inputs a single unit layer for the bona fide presentation (BP) vs. AP decision.

To optimise the network, we use the Binary Cross Entropy (BCE) loss, which is generally employed for binary classification tasks [15]. BCE $\mathcal{L}(\cdot)$ is computed as:

$$\mathcal{L}(x) = y \cdot \log p(x) + (1 - y) \cdot \log(1 - p(x)), \qquad (1)$$

where $p(x)$ is the predicted probability and y is the true label for the input x. We assign $y = 1$ for BPs and $y = 0$ for APs.

A. 1D audio waveforms to 2D spectrograms

Visualisation of audio/voice signals is key to many audio analysis tasks, often involving: i) time domain, ii) frequency domain, or iii) time and frequency domain representations known as spectrograms. These show the amplitude of the signal over time at a set of discrete frequencies. Recently, Gonzalez-Soler *et al.* [6] explored the feasibility of using several spectrogram-to-image strategies for voice PAD. As a result, the authors stated that image representations such as CQT [8] and STFT [9] can successfully preserve the features to distinguish a BP from an AP. Based on the above observations, we use those spectrogram-to-image representations in the first step of our algorithm:

- The STFT is a time-frequency decomposition based on the application of Fourier analysis to short-time segments or windows of the audio signal. In essence, it is effectively a filter bank where the bandwidth of each filter is constant and is related to the window function.
- The CQT is a perceptually motivated approach to time-frequency analysis. In this case, the bin frequencies of the filterbank are geometrically distributed. Compared to the STFT, the CQT has a greater frequency resolution for lower frequencies and a greater temporal resolution for higher frequencies.

B. Network architecture

As mentioned, our dual-stream temporal CNN comprises two streams: one optimised for the CQT representation and the other one for the STFT representation. The CQT is applied with a maximum frequency of $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 8kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \simeq 15$Hz (9 being the number of octaves). The number of bins per octave is set to 96. These parameters result in a time shift of 8.5ms. The STFT is implemented on a 30ms window with a 15ms shift and a 1024-point Fourier transform. For both STFT and CQT spectrogram-to-image representations, we perform a min-max normalisation and 8-bit quantisation on the log-magnitude spectrum. Based on this fact, we first split the input spectrograms into 5

TABLE I: General architecture of our proposed method.

Layers	CQT-stream	STFT-stream
Input	$5 \times 64 \times 128$	$5 \times 64 \times 128$
Backbone latent space	5×512	5×512
LSTM (4 layers)	1×256	1×256
Concatenation	1×512	
FC	1×256	
Sigmoid	1×1	

TABLE II: A summary of ASVspoof databases.

	Partition	Dataset	#BP samples	#AP samples
2019	LgA	training	2580	22800
		development	2548	22296
		evaluation	7355	63882
	PhA	training	5400	48600
		development	5400	24300
		evaluation	18090	116640
2021	LgA	evaluation	18452	163114

continuous frames, each of which has 64 $\times$128 pixels. A latent representation of 512 features per frame is computed using a given backbone (e.g., DenseNet, ResNet, or MobileNetv2). To exploit temporal information of speech images, the above latent representations are fed into 4 hidden LSTM layers each consisting of 256 neurons. The LSTM outputs of each stream are concatenated into a 512 vector, which in turn is further processed by a 256 FC layer. Finally, a fully connected layer of a single unit with sigmoid activation is added to produce the binary classification. A summary of the main architecture is shown in Tab. I.

In our implementation, we trained the network from scratch using the Adam optimiser [16]. A learning rate of 1×10^{-4} with a weight decay parameter of 1×10^{-6} was used. The framework was implemented in PyTorch [17] and trained on the Nvidia GPU Tesla M10 with 16 GB DRAM.

III. Experimental Setup

The experimental evaluation has a threefold goal: i) evaluate the detection performance of our proposed method over challenging scenarios, ii) analyse the effect of unbalanced data over the generalisation capabilities, and iii) establish a benchmark with the state-of-the-art PAD techniques. To that end, we focus on three different scenarios:

- *Known-attacks* includes an analysis of all PAI species. In all cases, PAI species for testing are also included in the training set, as described in [13].
- *Unknown PAI species*, in which the PAI species used for testing are not incorporated in the training set.
- *Cross-database* evaluates databases which are different in terms of PAI species, subjects, and capture devices from those used for training. We follow the cross-database protocol in [1], where the logical access evaluation partition in the ASVspoof 2021 is employed for testing.

A. Databases

The experimental evaluation is conducted over freely available databases ASVspoof 2019 [13] and 2021 [1] whose characteristics are summarised in Tab. II:

- ASVspoof 2019 database consists of two assessment scenarios: Logical Access (LgA) and Physical Access (PhA)[1]. Both LgA and PhA databases are partitioned into three disjoint datasets: training, development, and evaluation. The development set comprises samples created with the same set of PAI species as those in the training set (known-attacks evaluation). In contrast, samples in the evaluation set were generated with PAI species and capture conditions different to those in training (unknown-attack evaluation). In particular, the text-to-speech synthesis and voice conversion technologies used in the creation of the LgA evaluation samples are different from those used in the production of the training instances. In addition, configuration parameters such as reverberation levels, talker-to-ASV microphone distance, attacker-to-talker recording distance and loudspeaker qualities for the PhA evaluation samples are different from those used in the generation of the training PAIs [13].
- ASVspoof 2021 database includes an extra assessment scenario (i.e., DeepFake) along with the LgA and PhA scenarios. Following the protocol in [1], we use the LgA partition for the cross-database evaluation. In particular, the proposed algorithm is trained with LgA partition from the ASVspoof 2019 and evaluated over the same partition in the ASVspoof 2021.

[1]To avoid confusion with PA (presentation attack), we have named the two partitions of the ASVspoof 2019 database LgA and PhA

B. ISO Metrics for PAD

The experimental results are reported in compliance with the international standard ISO/IEC 30107-3 [14] for biometric PAD. We thus report the following metrics:

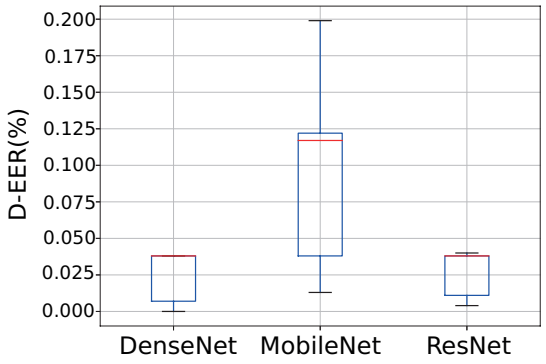
- Attack Presentation Classification Error Rate (APCER), which is the proportion of APs misclassified as BPs.
- Bona Fide Presentation Classification Error Rate (BPCER), which is the proportion of BPs misclassified as APs.

Based on the above metrics, we also report i) the Detection Error Trade-off (DET) curves between APCER and BPCER; ii) the BPCERs noted at different security thresholds (i.e., APCERs) such as 10% (BPCER10), 5% (BPCER20), and 1% (BPCER100), respectively; and iii) the Detection Equal Error Rate (D-EER), which is defined as the error rate value at the operating point where APCER = BPCER.

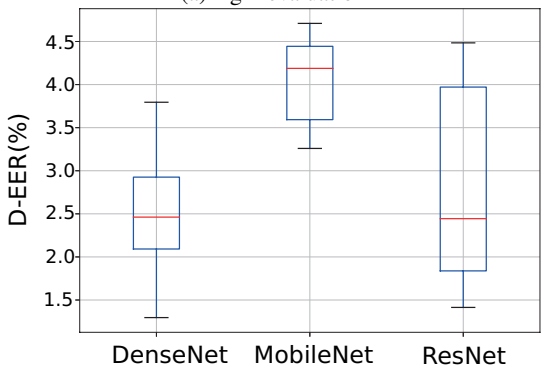
IV. Results and Discussion

A. Analysis on Known Attacks

In the first set of experiments, we evaluate the detection performance of our dual-stream temporal CNN for known

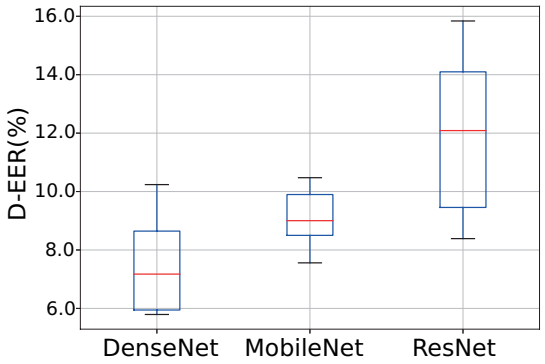


(a) LgA evaluation

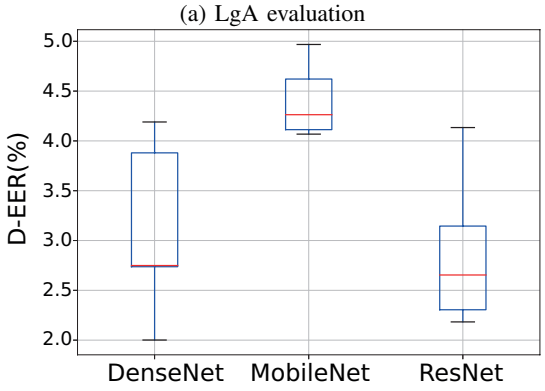


(b) PhA evaluation

Fig. 2: Detection performance per backbone for known attacks.



(a) LgA evaluation



(b) PhA evaluation

Fig. 3: Detection performance per backbone for unknown attacks.

attacks using three different backbones which have reported remarkable results in several pattern recognition tasks [18]: DenseNet with 121 layers [10], ResNet with 34 layers [11], and MobileNet version 2 [12]. As it can be seen in Tab. II, the AP samples represent 90% of the whole dataset. Therefore, we select randomly for this experiment the same number of AP samples as BPs at 5 different iterations. Then, we train our proposed method for each random subset and report the D-EER for known-attack scenarios in Fig. 2. As it can be observed, our approach achieves a mean D-EER lower than 0.10% and 4.04% for LgA and PhA, respectively. Whereas ResNet attains the best mean D-EER of 0.03% $\pm$ 0.02 for LgA, DenseNet reports the best mean D-EER of 2.51% $\pm$ 0.93 for PhA. However, we note that the minimum D-EER is yielded by DenseNet for LgA (i.e., D-EER = 0.00%) and PhA (i.e., D-EER = 1.30%). In addition, we observe that low standard deviations ranging 0.02-0.16 and 0.60-1.34 for LgA and PhA, respectively, indicate that the random selection of APs does not considerably impact the algorithm's detection performance.

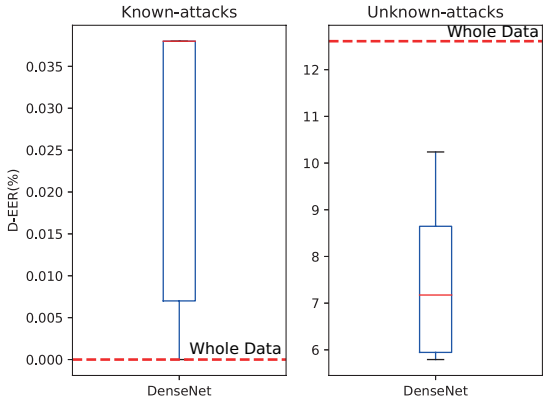
B. Analysis on Unknown Attacks

Now, we compute the detection performance of our approach combined with the three studied backbones (i.e., DenseNet, MobileNet, and ResNet) for the unknown-attack scenarios in Fig. 3. We observe that the mean D-EER is multiplied by a factor of 82 for DenseNet, 93 for MobileNet, and 457 for ResNet in comparison with the D-EERs reported for the known-attack evaluation. Specifically, the best performing backbone (i.e., DenseNet) achieves a mean D-EER of 7.56% $\pm$ 1.89, resulting in a minimum D-EER of 5.79%.

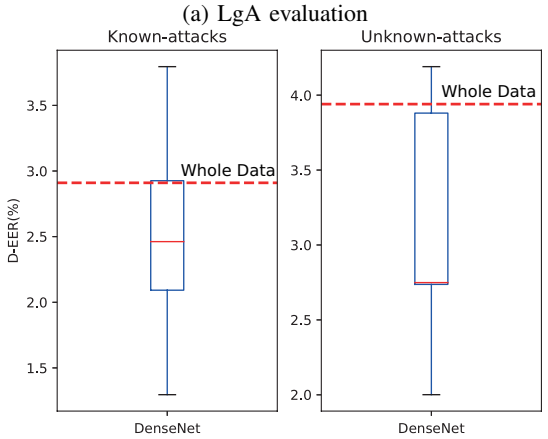
In contrast to the results reported for LgA, the detection performance per backbone for PhA is similar to those yielded for the known attacks. Mean D-EERs ranging 2.88%-4.41% confirm the observation done by [6] over the same experiment: the features for unknown samples stemming from the PhA partition follow a similar distribution as the ones for the spectrograms in the training set. Hence, we strongly think that algorithms for voice PAD should be able to achieve similar results for known and unknown attacks over the PhA partition.

C. Impact of Unbalanced Dataset over Unknown attacks

We evaluate to what extent the detection performance of our proposed method is affected when trained with the entire database. To that end, we selected the best performing backbone (i.e., DenseNet). In order to avoid bias in classifier training, we optimise the BCE loss in our approach by setting up weights per category (i.e., 0.90 for BPs and 0.10 for APs). Fig. 4 shows a benchmark of our proposed algorithm when it is trained with unbalanced (i.e., the entire dataset, red dashed line) and balanced (i.e., a random selection of AP samples) databases. We can see that the training with the entire database yields similar results to those attained by the random selection



(a) LgA evaluation



(b) PhA evaluation

Fig. 4: Benchmark of our proposed method trained with data random selection and the entire data (dashed red line).

of samples (see (a) and (b), column 1). Even, it achieves a D-EER of 0.00% for LgA (see a), column 1) which is lower than the mean value reported by training with a balanced database. This is because the features computed for the evaluation set follow the same distribution as those of the training set.

In contrast to the results reported for known attacks, we can observe that training with an unbalanced database considerably increases the D-EER compared to training using the same number of samples per category for unknown attacks. In particular, a D-EER of 12.61%, which is approximately twice higher than the one reported by the mean of the data random selection (i.e., 7.559%), is achieved for LgA. Subsequently, we can also note a decrease in the detection performance of our proposed approach when trained with the unbalanced database: a D-EER of 3.94% for an unbalanced database vs. a mean D-EER of 3.11% for a balanced database confirms the impact of training with an unbalanced database in the unknown-attack detection. We think that future studies focused on the PAD generalisation should consider the issues on unbalanced databases.

D. Cross-database evaluation

We evaluate the ability of our proposed method to spot PAIs across different databases. To that end, we follow the

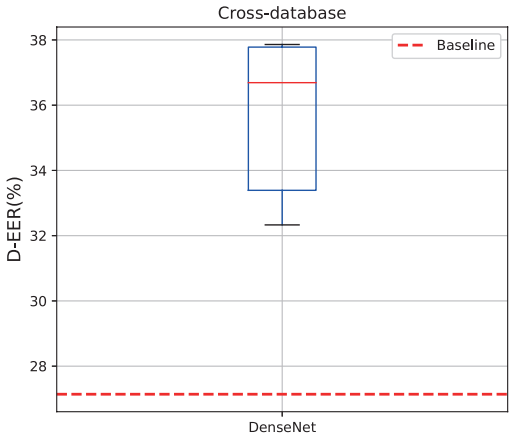
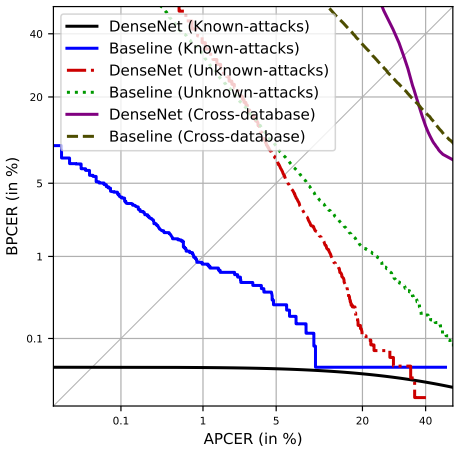


Fig. 5: Cross-database evaluation for the best performing backbone (i.e., DenseNet).

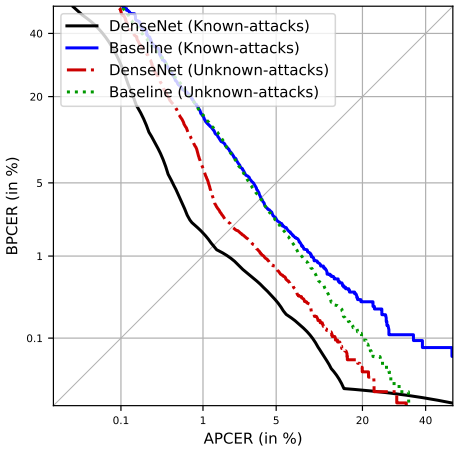
cross-database protocol defined in Sect. III and compute in Fig. 5 the D-EER for the best performing backbone (i.e., DenseNet) over the models trained over the five random sets mentioned in Sect. IV-A. We observe that the proposed method achieves a mean D-EER of 35.61% with a standard deviation of 2.58% which is higher than the one attained by the baseline [6] (i.e., 27.14%). Depending on the selection of the training set, a minimum D-EER of 32.33% is yielded, which shows that the selection of training samples is a challenge for PAD generalisation and should be taken into account in future research. In addition, these results confirm the need of enhancing the generalisation capability of neural networks. A considerable improvement of our results for this challenging scenario would be the combination of our dual-stream temporal CNN with those backbones which are developed for instance for domain adaptation [19]. Furthermore, the latest CNN families of EfficientNet architectures [20] could enhance the final decision of our framework.

E. Benchmark with the state-of-the-art

Finally, we benchmark in Fig. 6 the detection performance of our algorithm with the state-of-the-art for LgA and PhA. To that end, we select as a baseline the approach in [6] which reported a remarkable generalisation capability for the unknown attack detection. As it can be seen, our algorithm considerably outperforms the baseline for LgA and PhA. In particular, for the latter, a BPCER $\leq$ 1.78% and a BPCER $\leq$ 6.59% at an APCER $\geq$ 1.0% for known and unknown attacks, respectively confirm the soundness of our proposed method for operating over this challenging scenario. For LgA, the proposed techniques reports, for a high-security threshold (i.e., APCER $\geq$ 5.0%), a BPCER $\leq$ 0.07% and BPCER $\leq$ 7.74 for known and unknown attacks, respectively. Consistent with the results shown in Fig. 5, the cross-database performance computed by our framework for the best configuration in Fig. 5 (i.e., D-EER of 32.33%) suffers a decrease for high security thresholds, thus indicating the need for further research on this scenario.



(a) LgA evaluation



(b) PhA evaluation

Fig. 6: Benchmark with the state-of-the-art for known and unknown attacks and cross-database. Diagonal light-gray lines represent the D-EER (%).

V. Conclusions

In this paper, a dual-stream temporal CNN framework for the detection of AP attempts was proposed. The proposed method takes advantage of the temporal information provided by speech images which were previously obtained by transforming the spectrograms through different strategies (i.e., CQT and STFT). Those speech images are represented by a latent vector computed by any CNN. In our experiments, three traditional CNNs such as DenseNet, ResNet, and MobileNetv2 were tested, thus being DenseNet with 121 layers the best performing backbone for voice PAD. The experimental evaluation was conducted over the publicly available ASVspoof 2019 and 2021 in compliance with the metric defined in the international standard ISO/IEC 30107-3 for biometric PAD. The results showed that i) the training with an unbalanced database led to a detection performance decrease to spot unknown PAI species, ii) the random selection of APs for training the approach improved its generalisation capability, and iii) our proposed scheme was capable of outperforming the speech-image-based state-of-the-art methods for challenging scenarios where PAI species remain unknown in the training set (i.e., BPCER $\leq$ 6.59% for an APCER $\geq$ 1.0% over PhA). The decreasing detection performance of our algorithm for cross-database evaluation confirms the need for further research in the development of generalisable algorithms for this scenario.

References

[1] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[2] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[4] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey*, 2020, pp. 132–137.

[5] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "Stc antispoofing systems for the asvspoof2021 challenge," in *Proc. of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 61–67.

[6] L. J. Gonzalez-Soler, J. Patino, M. Gomez-Barrero, M. Todisco, C. Busch, and N. Evans, "Texture-based presentation attack detection for automatic speaker verification," in *Proc. Intl. Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–6.

[7] L. J. Gonzalez-Soler, M. Gomez-Barrero, and C. Busch, "On the generalisation capabilities of fisher vector based face presentation attack detection," *IET Biometrics*, vol. 10, no. 5, pp. 480–496, September 2021.

[8] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[9] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing (2Nd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[10] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[13] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *Proc. Interspeech*, 2019.

[14] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting*, International Organization for Standardization, 2017.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. NIPS-W*, 2017.

[18] K. Huang, A. Hussain, W. Wang, and R. Zhang, *Deep learning: fundamentals, theory and applications*, 2019, vol. 2.

[19] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Dual-adversarial domain adaptation for generalized replay attack detection," in *Proc. Interspeech*, 2020, pp. 1086–1090.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Intl. Conf. on Machine Learning*. PMLR, 2019, pp. 6105–6114.