IberSPEECH 2022
14-16 November 2022, Granada, Spain

Database dependence comparison in detection of physical access voice spoofing attacks

Manuel Chica Villar, Alejandro Gomez-Alanis, Eros Rosello, Angel M. Gomez, Antonio M. Peinado, Jose A. Gonzalez-Lopez

Department of Signals Theory, Telematics and Communications, University of Granada, Spain

manuelc@ugr.es, agomezalanis@ugr.es, erosrosello@ugr.es joseangl@ugr.es, amgg@ugr.es, amp@ugr.es

Abstract

The antispoofing challenges are designed to work on a single database, on which we can test our model. The automatic speaker verification spoofing and countermeasures (ASVspoof) [1] challenge series is a community-led initiative that aims to promote the consideration of spoofing and the development of countermeasures. In general, the idea of analyzing the databases individually has been the dominant approach but this could be rather misleading. This paper provides a study of the generalization capability of antispoofing systems based on neural networks by combining different databases for training and testing. We will try to give a broader vision of the advantages of grouping different datasets. We will delve into the "replay attacks" on physical data. This type of attack is one of the most difficult to detect since only a few minutes of audio samples are needed to impersonate the voice of a genuine speaker and gain access to the ASV system. To carry out this task, the ASV databases from *ASVspoof-challenge* [2], [3],[4] have been chosen and will be used to have a more concrete and accurate vision of them. We report results on these databases using different neural network architectures and set-ups.

Index Terms: Spoofing detection, Deep learning, Antispoofing, Speaker verification

1. Introduction

Speech is becoming a popular modality for human-computer interaction, thanks to recent advances in the fields of speech processing and deep learning. With the proliferation of voice biometric systems, one concern has recently attracted the attention of the research community: how to protect Automatic Speaker Verification (ASV) systems against impersonation attacks performed by malicious users claiming the identity of enrolled users. Thus, a malicious attacker could gain access to a system (e.g., a banking application) by claiming the identity of a genuine user by presenting voice samples of that user to the authentication system. There are many different types of spoofing attacks (a comprehensive review on this topic can be found in [5],[6]), but in this paper, we focus on replay attacks, as they can be performed even without technical expertise. In replay attacks (also known as physical access (PA) attacks),[7], [8], [9] as shown in Figure 1, the impersonator attempts to bypass the ASV system by presenting voice recordings of genuine users.

The spoofing attack proves to be one of the most difficult to detect, due to the fact that it can easily be executed, as only a small audio excerpt is required. In addition, the wide availability and widespread use of cell phones, recording devices, etc. facilitate the recording and playback of a genuine operator.

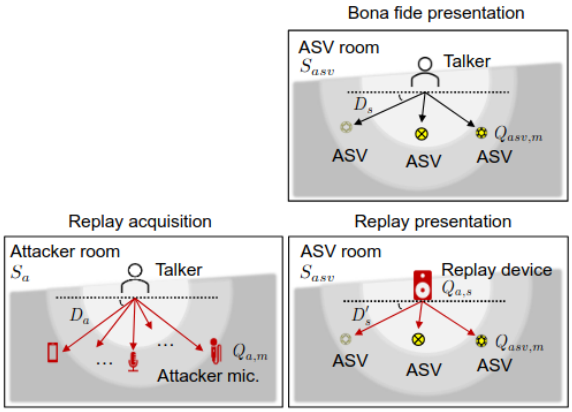

Figure 1: *Replay attack representation[4].*

Anti-spoofing is an identity theft detection technique used to prevent these impersonation attacks. However, the task of anti-spoofing is extremely difficult, as it can require a great deal of effort to distinguish the slightest differences between recorded and genuine versions.

Currently, the most common approach for the detection of spoofing attacks is based on the use of machine learning [10]. In this paper we consider two models, a Light Convolutional Neural Network (LCNN) and a baseline LFCC-LCNN system [8] operating upon Linear Frequency Cepstral Coefficients (LFCC) which are capable of analyzing voice recordings to determine whether they are bonafide or spoofed. The main problem we face is the need to obtain a large enough database to train our architecture.

The question arises as to how DNN models already trained can generalize when evaluated on other databases. The effectiveness of DNNs has been extensively demonstrated in terms of their generalization and deep learning capabilities, as can be seen in [11], [12], [13]. We will focus on evaluating the generalization capability of a state-of-the-art antispoofing model on unseen spoofing attacks. Our neural network performs a supervised learning process, with the model input being an audio signal and using input/output functions to consistently distinguish between authentic and spoofed audio.

To evaluate the capabilities of state-of-the-art, DNN-based systems to unknown spoofing attacks, we have chosen the *ASVspoof-challenge* databases that promote the design of countermeasures to protect automatic speaker verification systems. In fact, we will work with the datasets released in *2017*, *2019* and *2021* editions. We have also combined the 2017 and 2019 databases to determine the relationship between them and to

10.21437/IberSPEECH.2022-41

demonstrate whether significant results can be obtained that do not appertain with individual database training. Our models can generalize depending on the database used in the training and to check if joining databases can achieve a better result.

The structure of this paper is organized as follows. In section 2 we describe the elements used to carry out this task, as well as the database, neural network, and parameters used to evaluate our model. In section 3, we will show the results obtained and present an analysis and ideas that can be obtained from them. Finally, in section 4, we present a general conclusion and future works that can be derived from this work.

2. Methodology

2.1. Procedure

The general procedure followed has been to parameterize the training signals of an ASVspoof challenge dataset to train a DNN with them. Then, that DNN is evaluated with another test dataset to obtain a decision matrix. We will focus on physical access attacks. We will discuss and expose them according to [14], [15] and perform a treatment on them based on the study reported in [16]. To carry out the evaluation process, we followed the same procedure for the different datasets used in this paper (*ASVspoof_2017*, *ASVspoof_2019* and *ASVspoof_2021*). We evaluated two alternative spectral representations of the speech signals for anti-spoofing. First, we computed Short-Time Fourier Transform (STFT) features computed from 25 ms windows (using the Blackman window function) with 15 ms overlap. Also, Linear Frequency Cepstral Coefficients (LFCC) features were computed from 20 ms Hanning windows with 10 ms overlap.

2.2. Datasets

We evaluate the proposed anti-spoofing system on physical access (PA) attacks using the *ASVspoof challenge* databases from the 2017, 2019 and 2021 editions. This challenge is based on RedDots [17] project. We have chosen them because is a widely known challenge with a lot of information about the conditions under which the databases were created. They are authentic replay signals obtained in actual conditions changing the environment, using different microphones, speakers, and rooms. In the following, we provide more details about each database.

ASVspoof2017: The replay attacks of the *ASVspoof 2017* version were generated using 3 different quality categories (low, medium, high) of recording and playback devices. This dataset is designed to have a smaller number of recordings per session with more sessions of shorter duration in each one. One of the goals is to collect 52 sessions per speaker, one session each week, for one year. In this regard, each session is limited to two minutes. The composition of the sentences used for a recording session is shown in Table 1. This database was collected using mid- and, high-end smartphones and professional recorders (see [18] for more details). Spoofed utterances are the result of the replay and recording of authentic utterances using a variety of heterogeneous devices and acoustic environments. The latter is intended to simulate false utterance replay attacks [19] [20].

ASVspoof2019: The structure of the ASVSpoof physical-access database is summarized in Table 2. The dataset includes a total of 9 different replay configurations, comprising 3 categories of recording distances between the attacker and the speaker, and 3 categories of speaker quality. This scenario conforms as much as possible to the ISO definition of presentation attacks [21]. The 2019 edition is the first to focus on countermeasures for the three main types of attacks, i.e., those derived from text-to-speech (TTS), voice conversion (VC) and identity replay attacks. In addition, it is composed of simulated attacks with pre-recorded impulse responses. While the training and development sets contain spoofing attacks generated with the same algorithms/conditions (designated as known attacks), the evaluation data was generated with different randomly acoustic and replay configurations, (designated as unknown attacks).

ASVspoof2021:Finally, the ASVSpoof 2021 database contains only evaluation data for the PA task, i.e. no training data is provided with this database. For this task, the ASVspoof 2019 training data along with other external datasets are normally employed. The ASVspoof 2021 PA evaluation data comprises real bonafide and replayed samples similar to the ASVspoof 2017 database, but with a better-controlled design. Recordings are made in nine rooms in which three different types of microphones are placed at each of six different distances between the speaker and the ASV. Recordings are made with a total of 18 microphones simultaneously. Thus, there are 162 ($9 \times 3 \times 6$) different evaluation environments.

Table 1: *Structure of the ASVspoof 2017 physical access data corpus divided by training, development, and evaluation sets [3].*

Subset	Speaker	Replay Sessions	Replay Config	Utterances	
				Bonafide	Replay
Training	10	6	3	1507	1507
Development	8	10	10	760	950
Evaluation	24	161	57	1298	12008
Total	42	177	61	3565	14465

Table 2: *Structure of the ASVspoof 2019 physical access data corpus divided by training, development, and evaluation sets [2].*

	Speakers		Utterances	
Subset	Male	Female	Bonafide	Spoof
Training	8	12	5400	48600
Development	8	12	5400	24300
Evaluation	21	27	18090	116640
Total	37	51	28890	189540

2.3. AntiSpoofing Systems and Neural networks

The goal of anti spoofing systems is to avoid being fooled by spoofing attacks. This is done by maximizing the decision probability of the legitimate class and discarding the spoofing class. In this work, we evaluate two DNN-based front-ends for extracting embedding vectors [22] from the audio signals, in particular, two alternative implementations of the Light Convolutional Neural Network (LCNN), which has been shown to provide state-of-the-art results for anti-spoofing [23], were evaluated. These two architectures are described in detail in the following.

The architecture of the anti-spoofing system is shown in Figure 2. The baseline LFCC-LCNN system will be briefly described in section 2.4.1. A fragmentation using W frames

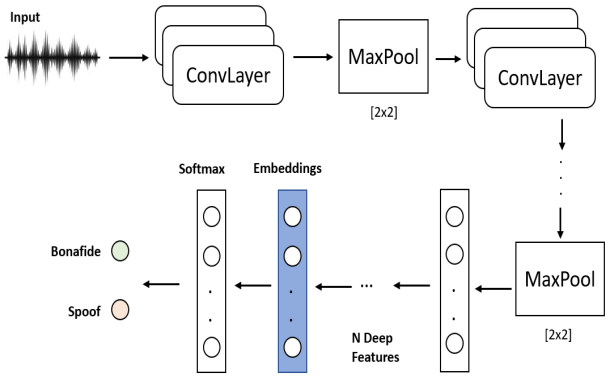

Figure 2: *Bonafide/Spoof detection and identity vector extraction (blue color). System 2*

is used to obtain the spectral features of the input signal that are fed into the system. Then, the CNN [24] provides one deep feature vector per window. The entire deep features vector of the considered utterance is processed by a state-of-art anti-spoofing system that was adapted from a previous work: a Light Convolutional Neural Network (LCNN), detailed in Section 2.4.2, which has shown to be very effective in detecting spoofed speech. This vector will be known as the "spoofing identity vector" and provides more discriminative information for spoofing detection than the raw speech features.

In this architecture the convolutional neural network acts as a frame-level deep feature extractor, providing a feature vector for each window of W frames. For this purpose, the CNN is trained to classify the input data as either bonafide or spoof.

After deep feature extraction, each utterance is represented by a single spoofing identity vector depicted in Figure 2 in blue. Finally, we use these vectors to make the final detection decision. To carry out this classification, we use Linear Discrimination Analysis (LDA). This metric assumes that each class can be modeled as a multivariate Gaussian as $N(\boldsymbol{x}|\boldsymbol{\mu}_k, \sum_k)$ naive. For a sample $\boldsymbol{x}$, the LDA model [25] uses the covariance $\sum_k$ and mean $\boldsymbol{\mu}_k$ of each class k and the dimension of the identity vector p. The goal of LDA is to find a linear transformation that maximizes the distance between classes while minimizing the dispersion within each class. In this case, LDA assigns a genuine speech confidence score to each utterance, which is then used for binary decision (spoof or genuine) during the evaluation.

2.3.1. System 1 - Baseline

A CNN is composed of multiple layers that enable the network to learn high-level abstract features from massive input data. Most CNNs have a deep multilayer structure with a large number of filter weights that increases the computational cost and the risk of overfitting. For this reason, we propose a Light Convolutional Neural Network, which can learn feature representations, even with a small number of training samples, and can achieve high accuracy due to its simple but sufficient modeling capability to learn deep features from speech inputs.

The first implementation is the one included with the ASspooch 2021 challenge. The baseline LFCC-LCNN [26] system operates upon Linear Frequency Cepstral Coefficients (LFCC) features feeding a LCNN. This system is applied to data with a maximum frequency of 8 kHz for the PA task using a 1024-point Fourier transform and 70 filters.

Table 3: *System 2 - LCNN architecture.*

LCCN			
Layer	Type	Filter/Stride	Output Channels
Layer 1	Conv	5x5/1x1	16
	MaxPool	2x2/2x2	8
Layer 2	Conv	1x1/1x1	16
	Conv	3x3/1x1	24
	MaxPool	2x2/2x2	12
Layer 3	Conv	1x1/1x1	32
	Conv	3x3/1x1	32
	MaxPool	2x2/2x2	16
Layer 4	Conv	1x1/1x1	32
	Conv	3x3/1x1	16
	MaxPool	2x2/2x2	8
Layer 5	Conv	1x1/1x1	16
	Conv	3x3/1x1	16
	MaxPool	2x2/2x2	8
-	FC1	-	128
-	FC2	-	2

The back-end is based on the LCNN reported in [27], but incorporates Long Short-Term Memory (LSTM) layers and average pooling. It is composed of 5 layers, combining convolution2d and normalization processes, then applying a 2x2 maxpooling matrix. Finally, we implement a 0.7 dropout operation used to reduce the risk of overfitting.
To perform the training process, we used a batch size of 64, a learning rate of $3{\cdot}10^{-5}$ with a decay of 0.5. A softmax activation function is applied to produce two-class predictions: bonafide or spoof.

2.3.2. System 2 - Light Convolutional Neural Network

The second network is an alternative implementation of the LCNN architecture, as depicted in Table 3, showing a summary of the LCNN architecture used. In this model, we apply T=400 frames and 864 filters to compute the STFT. The network consists of 5 layers, where each one has different light convolutional layers followed by a Maxpooling operation. This vector is then fed into a Fully Connected Layer (FC1) to obtain an utterances-level spoofing identity vector of 128 components. The proposed deep feature extractor was trained using the Adam optimizer with a learning rate of $3{\cdot}10^{-4}$ and a weight decay of 0.001. We used a batch size of 64 to train and evaluate. Also, batch normalization is applied to increase the stability and convergence of the training process. To avoid overfitting, a dropout of 0.7 was applied on the fully connected layer.

Finally, we evaluate using LDA, which provides us with a final score indicating whether the utterance is bonafide or spoofed.

2.3.3. Metrics

We used the Equal Error Rate (EER) [28], which is the point where the false acceptance rate and false rejection rate are equal, as the primary metric. We also report the results in terms of the minimum normalized tandem Detection Cost Function (t-DCF)[29]. This method extends the conventional DCF used in ASV research to scenarios involving spoofing attacks.

Table 4: *Decision matrix. ASVspoof Physical access evaluation scenarios in terms of EER (%) and t-DCF.*

		System 1-Baseline		System 2-LCNN	
Train	Eval	t-DCF	EER(%)	t-DCF	EER(%)
2017	2017	0.884	42.70	0.335	**12.83**
	2019	0.925	43.14	0.999	52.85
	2021	0.969	**38.47**	0.989	39.74
2019	2017	0.761	34.28	0.856	56.61
	2019	0.098	**3.76**	0.167	**6.30**
	2021	0.999	45.14	0.988	40.22
2017+2019	2017	0.614	26.21	0.851	32.71
	2019	0.141	**5.32**	0.150	**5.90**
	2021	0.997	48.21	0.999	42.12

Table 5: *Average EER values for the training and evaluation processes.*

		EER Measure (%)	
	Dataset	LCNN	LFCC-LCNN
Train	2017	35.09	43.08
	2019	34.36	27.12
	2017+2019	**26.95**	**24.75**
Eval	2017	34.03	34.43
	2019	**21.66**	**17.4**
	2021	40.67	43.93

3. Results

In this section, the performance of the tested approaches across the different databases introduced in the previous section is reported.

Table 4 shows the performance metrics obtained by both anti-spoofing systems as a function of the dataset used for training and the dataset used for evaluation. For training, we evaluated either using the training data included with the 2017 or 2019 databases alone or a combination of both (2017+2019). For evaluation, we only used the test sets defined for each database independently.

Due to changes introduced in the databases for the different ASV challenges, when we mix different sets of training and evaluation data corresponding to different years, the results we obtain are notably worse than those obtained by training and evaluating in the same year. Thus, we even reach an EER result of 52%. This result arises from performing training with the 2017 dataset and evaluation with the 2019 on System 2-LCNN. This can be explained because the 2017 dataset is the smallest dataset. Therefore, generalization is more difficult for our model.

The most remarkable result is obtained when we train our model by joining the 2017 and the 2019 datasets and evaluating with the 2019 dataset, where we get 5.9% in EER and 0.1497 in t-DCF. This means that by adding some different audio samples from the 2017 ASVspoof challenge to the 2019 ASVspoof challenge, we get better results than individually. Another result that proves this is obtained using the joint training (2017+2019) of the System 1 - Baseline network and evaluation in 2017, improving the results from 42.7% to 26%. In this case, the Anti-spoofing system seems to generalize better.

On the other hand, it can be seen how we do not obtain favorable results when we evaluate the 2021 dataset independently of the training dataset. Obtaining an EER between 38.47% and 48.21%. This can be explained by the fact that the evaluation set is too large compared to the training and development set, which implies that our model does not fit and generalize correctly.

In Table 5 we can see the average EER results for each dataset. This allows us to observe which database gives better results in training and validation independently. That is, we can observe which training set performs best for training a model independently of the set with which it is to be evaluated later and which set is best for evaluating, independently of which database a model has been trained with.

In order to analyze which dataset is better for training, the mean has been calculated over the EER results from the 3 evaluations (2017, 2019 and 2021) obtained from the same training set. In the case of evaluation, the mean has been calculated with the EER of each training dataset (2017, 2019 and 2017+2019), all evaluated from the same dataset.

By comparing EERs, it can be seen that the best results are obtained by training both systems with the data from the joint databases, (2017 + 2019). On the other hand, it is observed that the worst results appear when we evaluate using the 2021 dataset. This may be because the 2021 evaluation set contains 45 Giga bytes of audio, which is more than twice as much as each training set. A possible solution to improve this result would be to allocate a higher percentage of audio from the full set to the training and development process so that our model can perform better generalization.

4. Conclusions and future work

In this paper, we have studied the relationship between different datasets used in three well-known ASVspoof challenge series, showing that there may be relevant results that could be being ignored.

We have highlighted the limitations of recently proposed databases for anti-spoofing challenges in assessing the actual ability of DNN networks to generalize with new data. Thus, anti-spoofing solutions with reasonably high EER scores may naturally fail when presented with unseen spoofing attacks.

In future work it would be worthwhile to investigate the result of combining a larger number of databases, coming from other challenges. Other anti-spoofing systems such as the Gated Recurrent Convolutional Neural Network (GRCNN) [30] should be tested in order to check their generalization capabilities with unseen attacks.

5. Acknowledgements

This work has been supported by the FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades Proyecto PY20 00902 and by the project PID 2019-104206GB-IOO funded by MCIN/AEU/10.13039/501100011033.

6. References

[1] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, 2021.

[2] 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," *arXiv preprint arXiv:1904.05441*, 2019.

[3] 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. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 296–303.

[4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, A. N. Jose Patino, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *Proc. Interspeech, 2021*, 2021.

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H.Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun*, vol. 66, pp. 130–153, Feb. 2015.

[6] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, and S. King, "Anti-spoofing for text.independent speaker verification: An initial database comparison of countermeasures and human perfomance," *IEE/ACM Trans. Audio, Speech, Lang. Preocess.*, vol. 24, pp. 768–783, 2016.

[7] D. R. Campbell, K. J. Palomaki, and G. Brow, ""a matlab simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems Journal, ISSN*, vol. 9, pp. 1352–9404, 2016.

[8] E. Vincent., "http://homepages.loria.fr/evincent/software/roomsimove 1.4.zip," *Roomsimove*, 2008.

[9] A. Novak, P. Lotton, and L. Simon, "Synchronized swept-sine: Theory, application, and implementation," *J. Audio Eng. Soc*, vol. 9, pp. 786–798, 2016.

[10] L'Heureux, Alexandra, Grolinger, Katarina, Elyamany, H. F., Capretz, and M. A. M., "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[11] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with dnn and one-class svm for the asvspoof 2015 challenge," *Proc. Interspeech, 2015*, 2015.

[12] C. Zhang, "Joint information from nonlinear and linear features for spoofing detection: an i-vector/dnn based approach," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process*, 2016.

[13] N. Chen, Y. Qian, H. Dinkel, B. Chen, , and K. Yu, "Robust deep feature for spoofing detection," *The SJTU system for ASVspoof 2015 challenge*, 2015.

[14] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudan-pur, "A study on data augmentation of reverberant speech for ro-bust speech recognition," *Proc. IEEE Int. Conf. on Acoustics,Speech and Signal Processing*, pp. 5220–5224, 2017.

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Proc. IEEE Int. Conf. on Acoustics*, pp. 5329—-5333, 2018.

[16] A. Janicki, F. Alegre, and N. Evans, "In assessment of automatic speaker verification vulnerabilities to replay spoofing attacks," *Security and Communication Networks*, vol. 9, pp. 3030–3044, 2016.

[17] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. B. ummer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, M. J. Alam, A. Swart, and J. Perez, "The reddots data collection for speaker recognition," *Proc. INTERSPEECH*, pp. 2996–3000, 2015.

[18] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautam, D. Thomsen, A. Sarkar, Z. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautam, and K. A. Lee, "Reddots replayed: A new replay spoof-ing attack corpus for text-dependent speaker verification research," *Proc. ICASSP,New Orleans, USA*, 2017.

[19] Z. Wu, S. Gao, E. Chng, , and H. Li, "A study on re-play attack and anti-spoofing for text-dependent speaker verification," *Proc. ICASSP,New Orleans, USA*, pp. 1–5, 2014.

[20] Alegre, A. Janicki, , and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," *Proc. ICASSP,New Orleans, USA*, pp. 157–168, 2014.

[21] I. 30107, "Biometric presentation attack detection," *Information technology*, 2016.

[22] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. Magimai.-Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2021.

[23] J. G. A. Gomez-Alanis, A.M. Peinado, "Robust speaker verification system based on deep neural network," *Department of Signal Theory Telematics and Communications*, 2022.

[24] Albawi, Saad, Mohammed, T. Abed, Al-Zawi, and Saad, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.

[25] A. Gomez-Alanis, A. Peinado, J. Gonzalez Lopez, and A. Gomez, "Performance evaluation of front- and back-end techniques for asv spoofing detection systems based on deep features," 11 2018, pp. 45–49.

[26] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," 2022. [Online]. Available: https://arxiv.org/abs/2210.02437

[27] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *Proc. Interspeech*, 2019.

[28] N. Brummer and E. de Villiers, "Theory, algorithms and code for binary classifier score pro-cessing," *The BOSARIS toolkit*, 2011.

[29] T. Kinnunen, K.-A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D.-A. Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," *Proc. Odyssey*, 2018.

[30] A. Gómez Alanís, J. A. González López, and A. M. Peinado Herreros, "Ganba: Generative adversarial network for biometric antispoofing," *MDPI*, 2022.