Investigating Light-ResNet Architecture for Spoofing Detection under Mismatched Conditions

Prasanth Parasu[1], *Julien Epps*[1], *Kaavya Sriskandaraja*[1], *Gajan Suthokumar*[1]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia

prasanth.parasu@unsw.edu.au, j.epps@unsw.edu.au, k.sriskandaraja@unsw.edu.au, g.suthokumar@unsw.edu.au

Abstract

Current approaches to Voice Presentation Attack (VPA) detection have largely focused on spoofing detection within a single database and/or attack type. However, for practical Presentation Attack Detection (PAD) systems to be adopted by industry, they must be able to generalise to detect diverse and previously unseen VPAs. Inspired by successful aspects of deep learning systems for image classification such as the introduction of residual mappings through shortcut connections, this paper proposes a novel Light-ResNet architecture that provides good generalisation across databases and attack types. The introduction of skip connections within residual modules enables the training of deeper spoofing classifiers that can leverage more useful discriminative information learned in the hidden layers, while still generalising well under mismatched conditions. Utilising the wide variety of databases available for VPA research, this paper also proposes a set of generalisation evaluations which a practical PAD system should be able to meet: generalising within a database, generalising across databases within attack type and generalising across all spoofing classes. Evaluations on the ASVspoof 2015, BTAS 2016 (replay) and ASVspoof 2017 V2.0 databases show that the proposed Light-ResNet architecture is able to generalise across these diverse tasks consistently, outperforming CQCC-GMM and Attentive Filtering Network (AFN) baselines.

Index Terms: spoofing detection, speaker verification, residual neural network, generalisation

1. Introduction

Biometric authentication has been gaining popularity as a form of identification and access control for many systems. In particular, Automatic Speaker Verification (ASV) systems are being used to protect secure or sensitive facilities such as office buildings and for authentication in telephone banking [1]. However, a major deterrent preventing the adoption of ASV is its susceptibility to malicious spoofing attacks, otherwise known as Voice Presentation Attacks (VPAs). Spoofing attacks aim to impersonate a valid user and fall into four classes: Impersonation, Speech Synthesis (SS), Voice Conversion (VC) and replay.

A convenient method to deter a VPA is through anti-spoofing or presentation attack detection (PAD) systems, which perform statistical acoustic characterisation of genuine and spoofed speech. However, a key challenge is their inability to generalise to detecting new kinds of VPAs. Many PAD systems fail to detect unseen SS or VC attacks or are unable to detect a replay attack from a different recording session [2]. Many systems also use prior knowledge about the development of the spoofing technique to detect and deflect attacks, which is not suitable in practical scenarios where knowledge about a specific spoofing technique will not be known ahead of time [3]. State-of-the-art PAD systems have usually been developed to tackle either SS/VC or replay attacks, not both, and thus fail to generalise across spoofing classes. To solve these problems, many researchers have turned to score fusion techniques to combine the outputs of several PAD systems, however these may be impractical outside of controlled research challenges [4, 5].

To date, data-centric challenges such as ASVspoof 2015 and 2017 have focused solely on detecting spoofing attacks of the same type. Given the deficiencies of current systems, a number of papers showed a performance degradation when specific features optimised for one database were tested on another database [3, 6, 7, 4, 8, 9]. However, many studies have limited their scope to cross-corpus evaluation within the same spoofing class and have not explored developing a PAD system to mitigate SS, VC and replay attacks.

Deep learning systems developed initially for problems from image classification to natural language processing have been successfully applied to VPA detection. For example, the best-performing system in ASVspoof 2017 utilised a Light Convolutional Neural Network architecture to discriminate between genuine and replay attack utterances [10]. Similarly, a feedforward Deep Neural Network classifier demonstrated impressive results detecting SS/VC attacks [11]. Overall, although many systems demonstrate great intra-database generalisation, their ability to generalise in cross-corpus scenarios or against SS/VC and replay attacks remains unknown.

Deep residual networks (ResNets) in particular were introduced to address the neural network degradation problem, where it was found that adding more layers to a network did not necessarily improve its generalisation ability from training set to test set for the image classification problem [12]. The skip connections in residual modules do show promise for relieving the degradation problem, and the ResNet architecture has demonstrated impressive generalisation for image recognition in [12], and a number of promising PAD systems such as the GD-ResNet-18 with Attention [13] and Attentive Filtering Network (AFN) [14] also used residual modules. However, pure ResNet architectures have not previously been shown to be effective in combating replay attacks such as in [15, 16]. Furthermore they have yet to be evaluated on SS/VC attacks.

Herein, we propose a novel Light-ResNet architecture with spectrogram input features and investigate its generalisation. Following the above motivation to mitigate SS, VC and replay attacks, we evaluate using three criteria: ability to generalise in intra-database evaluation (i.e. between training, development and evaluation data sets inside the same database), ability to generalise in cross-corpus evaluation (i.e. trained on one database and evaluated on another) and finally ability to generalise across spoofing classes (i.e. system trained on three classes SS, VC and replay can be used to detect all three).

2. Light-ResNet Architecture

2.1. Residual Module Architecture

2.1.1. Addressing the neural network degradation problem

The network degradation problem hinders the ability of deep-learning based PAD systems to generalise to new and diverse spoofing attacks, as new layers make discriminating between genuine and spoofed utterances more difficult. Residual modules address this problem through the introduction of skip connections as seen in Figure 1, allowing the network to learn the identity function easily. Consider adding an extra layer to a neural network: if the new layer isn't helpful for discriminating between genuine and spoofed utterances, the layer should ideally have less influence on the classification decision (i.e. by learning an identity mapping). Achieving an identity mapping with a stack of nonlinear layers is more difficult than simply pushing the weights of the residual module to zero [12].

2.1.2. Addressing the gradient vanishing problem

The gradient vanishing problem is also addressed through skip connections. We can imagine the skip connections as helping preserve the gradient of the error during backpropagation. As we backpropagate through the skip connection, the error gradient will simply be multiplied by 1. This allows us to train much deeper networks, enabling our PAD system to learn more discriminating features between genuine and spoofed speech.

2.1.3. Post-activation vs pre-activation

An improvement to the standard fundamental residual module was proposed in [17] for image classification and is yet to be explored in the context of VPA detection with convolutional layers. In the new residual module named "pre-activation", the Batch Normalisation and ReLU layers are moved to before the convolutional layer as highlighted in Figure 1. Our preliminary experimental work found that the "pre-activation" module provided to a reduction in overfitting relative to post-activation. This is linked to the batch normalisation regularisation effect [17]. In the "pre-activation" module, the inputs to all convolutional layers are normalised.

2.2. Light-ResNet Architecture

The results from our own preliminary work discussed in Section 4.1 demonstrate that ResNet systems previously developed for replay attack detection in [15, 16] perform poorly due to lack of appropriate parameter selection and tuning, as well as improper input feature selection. The ResNet architectures originally proposed for image classification in [12] were trained on much larger datasets than are available for VPA detection. Thus applying the default configurations (i.e. number of filters, number of layers) designed for image classification would not be appropriate for developing a generalised spoofing detector.

It is imperative for the problem of spoofing detection that a reduced version of the ResNet architecture be adopted. Reduced parameters aid in preventing overfitting and can be achieved with a reduction in the number of filters, layers and smaller filter size. Figure 1 shows an 18 layer ResNet as each residual block contains two residual modules, however the number of residual modules in each layer can vary, i.e. 34 layer ResNet has configuration [3, 4, 6, 3], number of fundamental residual modules in each residual block.

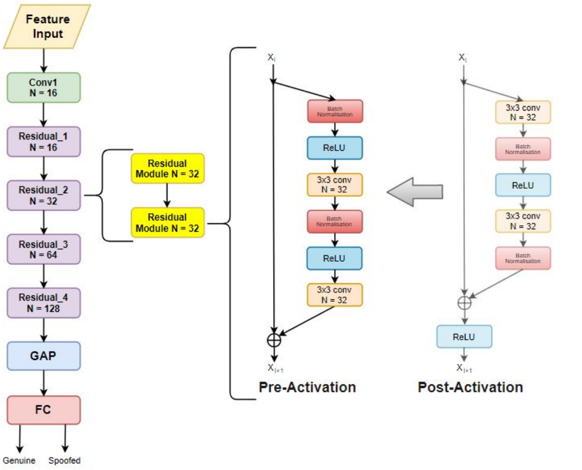

Figure 1: *Proposed deep Light-ResNet architecture, developed based on image classifier in [12]. Each residual block (i.e. Residual_M) is made up of a number of fundamental residual modules. N represents the number of filters in each convolution layer. Skip connections help preserve the gradient of the error during backpropagation and help learn the identity function easily, while replacing post-activation (conventional) with pre-activation modules mitigates overfitting.*

We chose to use spectrograms as the input features as they have shown success in other deep learning PAD systems, as seen in [14, 18] containing useful discriminative information that convolutional neural networks can learn patterns from. Furthermore authors in [18] highlighted how CNN classifiers were able to extract more discriminative information from simple spectrograms compared with more complex features such as Perceptual Minimum Variance Distortionless Response (PMVDR). These spectrogram features are fed to a Light-ResNet classifier which determines whether the utterance is genuine or spoofed.

3. Experimental Settings

3.1. Spectrogram Feature

Spectrograms were created using Python with the SciPy [19] library, with a hanning window length of 25ms, frame shift of 10ms and 512 DFT points. Cepstral mean and variance normalization (CMVN) was applied using the scikit-learn library [20]. An utterance length of 5s was chosen, with longer utterances being truncated and shorter utterances being repeated such that all utterances are of the same length. The largest database investigated (ASVspoof 2015) contains utterances with lengths in the order of 1-2s [21], thus 5s was chosen to prevent excessive repetition, although it means we need to truncate longer utterances such as in BTAS 2016.

3.2. ResNet Backend

The backend first consists of a 7x7 convolution layer and a 3x3 max pool layer with strides of 2, in order to down sample the input tensor (Conv1). This is followed by four residual blocks, as seen in Figure 1 detailing the kernel sizes and number of filters in the ResNet-18 architecture. A stride of 1 was used in all convolutional layers except the first convolutional layer, where a stride of 2 was used, to further down-sample the input tensor. This is followed by a global average pooling layer (GAP), so that we only have a single value per feature map. This is then passed to a fully connected layer with the sigmoid activation function, classifying the utterance. An l2 weight regulariser also known as weight decay of 0.0001 was also applied to all convolutional layers. Note that many of the chosen parameters were based on the original work in [12], while the two key parameters, number of filters in each layer and number of layers are further investigated in Section 4.

During training the Adam optimiser was used with a learning rate of 0.001 and trained for 15 epochs. Equal Error Rate (EER) was used as the metric to evaluate the performance of the PAD system. The scikit-learn Python library was used to calculate the EER [20].

3.3. Databases and Evaluation Methodology

The proposed system was evaluated on three databases: ASVspoof 2015, 2017 V2.0 and BTAS 2016 (replay). The ASVspoof 2015 dataset was used to evaluate the performance of the proposed systems on genuine, SS and VC attack utterances. The ASVspoof 2017 V2.0 dataset was used to evaluate the performance of the proposed systems on genuine and replay attack utterances. Finally the BTAS 2016 dataset was introduced to investigate the performance of the proposed systems in cross-corpus scenarios. The BTAS 2016 database contains genuine utterances, SS, VC and replay attacks, however only the genuine and replay attack utterances of the BTAS 2016 dataset were investigated in this paper.

Table 1: *Number of utterances, in each section of the ASVspoof 2015, BTAS 2016 (replay) and ASVspoof 2017 V2.0 databases*

Database	Number of utterances		
	Training	Dev.	Eval.
ASVspoof 2015 [21]	16376	53372	193404
BTAS 2016 (replay) [22]	7773	7795	10376
ASVspoof 2017 [23]	3014	1710	13306

The proposed PAD system is evaluated against the above databases in the following way:

- Generalise in intra-database evaluation: Training the proposed system on each of the ASVspoof 2015, 2017 V2.0 and BTAS 2016 (replay) training sets and evaluating on their corresponding evaluation dataset.
- Generalise in cross-corpora evaluation (replay attacks): Training the proposed system on the ASVspoof 2017 V2.0 training set and evaluating on evaluation section of BTAS 2016 (replay) and vice versa.
- Generalise in unified-spoofing evaluation (i.e. across spoofing classes): Training the proposed system on the training sets of the ASVspoof 2015 and 2017 V2.0 databases and evaluating the system on the evaluation section of both databases.

Note that whenever we train with ASVspoof 2017 V2.0, our training data pools utterances from the training and development set.

3.4. Baselines

The first baseline explored consists of a CQCC features extracted in the frontend. For SS, VC attacks standalone delta-delta coefficients have shown to be effective [24], while retaining all the coefficients (static, delta and delta-delta coefficients) is beneficial for replay detection [25]. As we are investigating generalised spoofing detection, we retain all coefficients. This is fed to a 2-class GMM that classifies the sample as either genuine or spoofed utterance. We used the same implementation as the baseline configuration for CQCC-GMM in [26].

The AFN was proposed in [14] for replay detection on the ASVspoof 2017 V2.0 database. Our implementation of the AFN differs in that shorter utterances (5s) are used to train and evaluate this system to be comparable to the proposed system. Furthermore shorter utterances are more suitable for evaluation particularly on the ASVspoof 2015 database.

4. Results and Discussion

4.1. Investigating the ResNet Architecture

We first develop the ResNet classifier architecture for spoofing detection specifically, as the parameters provided for image classification in [12] may not necessarily translate to VPA detection. The tuning was conducted on ASVspoof 2017 V2.0 database. Note for this Section we train only on the training set (i.e. development set not included).

4.1.1. Number of filters

In Figure 2, we plot the Evaluation EER (%) for the different number of filters used in convolutional layers in each residual block in Figure 1. The general trend that we notice is that as we reduce the number of filters in the network the generalisation of the ResNet improves. A filter configuration of [16, 32, 64, 128] was found to be optimal, decreasing the total amount of parameters for a 18 layer ResNet from 11,182,082 (original ResNet in [12] with [64, 128, 256, 512] filters) to 701,538. The reduction in number of parameters allowed the network to generalise to detecting previously unseen replay attacks.

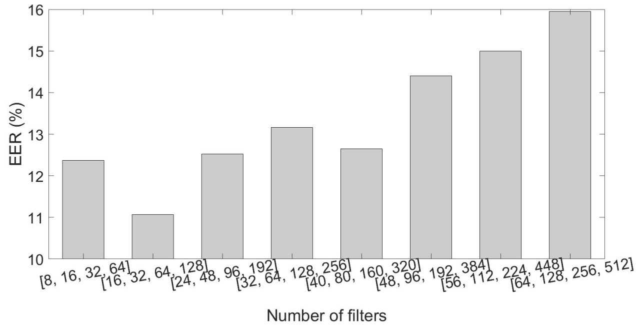

Figure 2: *EER vs filter configuration; General trend captured highlights how reduction in number of parameters leads to improved generalisation.*

4.1.2. Number of layers

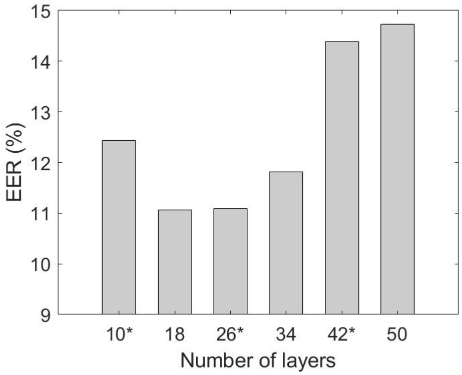

Figure 3: *EER vs Number of Layers; layers with * represent systems not proposed for image classification in [12], but investigated here to understand the effect; tradeoff between deeper network learning new discriminative information and number of network parameters.*

Table 2: *EER (%) for baseline and proposed systems for intra-database evaluation*

System	Intra-database Evaluation			
	ASVspoof 2017	ASVspoof 2015		BTAS 2016 (replay)
	Eval EER (%)	S10 EER (%)	Eval EER (%)	Eval EER (%)
CQCC-GMM [23, 27, 28]	24.77	13.407	2.83	8.32
AFN [14]	11.16	1.38	1.35	**3.07**
ResNet-34	12.81	1.39	0.33	8.00
Light-ResNet-18	9.73	1.14	0.35	4.88
Light-ResNet-34	**9.67**	**0.81**	**0.19**	5.55

Table 3: *EER (%) for baseline and proposed systems for cross-corpus and unified spoofing evaluation; For the cross-corpus scenario when evaluating on ASVspoof 2017, training was performed on BTAS 2016 (Replay) and vice versa*

System	Cross-corpus Evaluation		Unified Evaluation			Intra, Cross
	ASVspoof 2017	BTAS 2016 (replay)	ASVspoof 2017	ASVspoof 2015		& Unified
	Eval EER (%)	Eval EER (%)	Eval EER (%)	S10 EER (%)	Eval EER	Aggregated Rank
CQCC-GMM	39.23	28.78	21.42	18.86	3.96	5
AFN	16.01	15.52	**13.50**	15.72	3.44	2
ResNet-34	27.06	19.46	25.91	10.59	**2.18**	4
Light-ResNet-18	21.74	18.63	15.77	15.40	3.91	3
Light-ResNet-34	**15.21**	**14.55**	16.38	**9.28**	2.99	1

Next we can investigate the effect of the number of layers on generalisation within the ASVspoof 2017 V2.0 database as seen in Figure 3. We find that the 18 layer ResNet produces the best results when the number of filters have the fixed configuration [16, 32, 64, 128]. Furthermore as we increase the number of layers, the generalisation of the ResNet system may not improve. Thus there is a tradeoff between new discriminatory information learnt in additional hidden layers and overfitting to the training dataset due to additional parameters.

4.2. Intra-database Evaluation

We first investigate training and evaluation within each database. We follow the evaluation methodology set out in 3.3. We see that the Light-ResNet-34 architecture outperforms the other systems on the ASVspoof 2017 V2.0 database and ASVspoof 2015 database. It is interesting to note that the ResNet-34, with filter parameters [64, 128, 256, 512], performs poorly on ASVspoof 2017 and BTAS 2016 (replay) compared to the other deep learning based classifiers, however performs well on ASVspoof 2015. We can attribute this to the size of the ASVspoof 2015 database being much larger than the other databases, thus doesn't overfit to the training data even with more parameters. It is also clear that the ResNet-34 overfits to the training set across all databases compared to the Light-ResNet-34, leading to poorer performance.

4.3. Cross-Corpus and Unified-spoofing Evaluation

Investigating the Cross-Corpus condition, training on BTAS 2016 (replay) and evaluation on ASVspoof 2017 V2.0 and vice versa, we first find that the deep learning based architectures generalise better than the CQCC-GMM baseline. Next comparing the Light-ResNet-34 with ResNet-34 we see the performance improvement that can be obtained by a reduced network configuration. Another interesting comparison is the superior performance of Light-ResNet-34 over Light-ResNet-18, which can be attributed to additional hidden layers in the ResNet learning useful information, allowing the classifier to better discriminate between genuine and spoofed speech. This lines up with our theoretical understanding developed in Section 2, with skip connections preventing the performance of deeper networks from degrading if they are unable to learn useful discriminatory information. Similarly, the Light-ResNet-34's ability to outperform the AFN can be attributed to its deeper architecture, despite the Light-ResNet-34 having 7 times as many parameters.

Next looking at the unified evaluation condition, we see that the performance of all systems degrades when SS, VC and replay attacks are combined in the training set. The proposed Light-ResNet-18 and 34 are outperformed by the AFN when evaluating on replay attacks. However the ResNet-34 and Light-ResNet-34 perform particularly well on SS/VC attacks. This is especially evident on the S10 attack which doesn't use a vocoder for waveform generation and thus the artefacts introduced are different to those seen in the training set. The performance of the Light-ResNet-34 on the S10 attack, highlight its ability to generalise to new, previously unseen VPAs. The lack of performance degradation on the ASVspoof 2015 dataset between the Light-ResNet-34 and ResNet-34 can be attributed to the large amount of SS/VC training examples present unlike replay attacks.

Overall, aggregating the results of the systems using Borda's method we find that the proposed Light-ResNet-34 system produces the best performance across the intra, cross-corpus and unified conditions.

5. Conclusion

In this paper we have investigated the generalisation performance of the proposed Light-ResNet system under various conditions including within databases, across databases within attack type and across spoofing classes. This showed that Light-ResNet-34 outperformed the CQCC-GMM and the AFN across these diverse mismatched conditions, which we attribute to the architecture's capacity for deep hidden layers to learn discriminative spoofing-related information without overfitting. In particular, the skip connections address the gradient vanishing problem and allow for deeper networks to be trained. Light-ResNet-34's EER for the S10 attack in the ASVspoof 2015 database is especially indicative of its generalisation performance, given that the S10 attack is completely unknown. In future, we plan to investigate the systems in this paper without truncating file lengths and investigate the cross-corpus performance of the proposed systems for SS/VC attacks.

6. References

[1] J. Kollewe, “Hsbc rolls out voice and touch id security for bank customers,” *The Guardian*, 2016. [Online]. Available: https://www.theguardian.com/business/2016/feb/19/hsbc-rolls-out-voice-touch-id-security-bank-customers

[2] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, *Introduction to Voice Presentation Attack Detection and Recent Advances*. Springer, 01 2019, ch. 15, pp. 321 – 361.

[3] D. Paul, M. Sahidullah, and G. Saha, “Generalization of spoofing countermeasures: A case study with asvspoof 2015 and btas 2016 corpora,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2047–2051.

[4] P. Korshunov and S. Marcel, “Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 695–705, June 2017.

[5] A. R. Goncalves, R. P. V. Violato, P. Korshunov, S. Marcel, and F. O. Simoes, “On the generalization of fused systems in voice presentation attack detection,” in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2017, pp. 1–5.

[6] M. Todisco, H. Delgado, and N. Evans, “Constant Q cepstral coefficients: A spoofing countermeasure for automatic Speaker verification,” *Computer Speech & Language, 20 February 2017*, 02 2017. [Online]. Available: http://www.eurecom.fr/publication/5146

[7] P. Korshunov and S. Marcel, “Cross-database evaluation of audio-based spoofing detection systems,” in *Interspeech*, 09 2016.

[8] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, “Cross-domain replay spoofing attack detection using domain adversarial training,” in *INTERSPEECH 2019*, 2019.

[9] P. Korshunov and S. Marcel, *A cross-database study of voice presentation attack detection*. Springer, 01 2019, ch. 16, pp. 363–389.

[10] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, K. Oleg, and V. Shchemelinin, “Audio replay attack detection with deep learning frameworks,” in *Interspeech*, 08 2017, pp. 82–86.

[11] H. Yu, Z. Tan, Z. Ma, R. Martin, and J. Guo, “Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4633–4644, Oct 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[13] F. Tom, M. Jain, and P. Dey, “End-to-end audio replay attack detection using deep convolutional networks with attention,” in *Interspeech*, 2018.

[14] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, “Attentive filtering networks for audio replay attack detection,” *CoRR*, vol. abs/1810.13048, 2018.

[15] Z. Chen, Z. Xie, W. Zhang, and X. Xu, “Resnet and model fusion for automatic spoofing detection,” in *Interspeech*, 2017.

[16] W. Cai, C. Danwei, W. Liu, G. Li, and M. Li, “Countermeasures for automatic speaker verification replay spoofing attack : On data augmentation, feature representation, classification and fusion,” in *Interspeech*, 08 2017, pp. 17–21.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” *ArXiv*, vol. abs/1603.05027, 2016.

[18] C. Zhang, C. Yu, and J. H. L. Hansen, “An investigation of deep-learning frameworks for speaker verification antispoofing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, June 2017.

[19] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, “SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python,” *arXiv e-prints*, p. arXiv:1907.10121, Jul 2019.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, and M. Sahidullah, “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” in *Interspeech*, 09 2015.

[22] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. G. S. Mello, R. P. V. Violato, F. O. Simões, M. U. Neto, M. A. Angeloni, J. A. Stuchi, H. Dinkel, N. Chen, Y. Qian, D. Paul, G. Saha, and M. Sahidullah, “Overview of btas 2016 speaker anti-spoofing competition,” *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6, 2016.

[23] H. Delgado, M. Todisco, M. Sahidullah, and N. W. D. Evans, “Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements,” in *Odyssey*, 2018.

[24] M. Sahidullah, T. Kinnunen, and C. Hanilçi, “A comparison of features for synthetic speech detection,” in *INTERSPEECH*, 2015.

[25] R. Font, J. Espín López, and M. Cano, “Experimental analysis of features for replay attack detection — results on the asvspoof 2017 challenge,” in *Interspeech*, 08 2017, pp. 7–11.

[26] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, “An investigation about the scalability of the spoofing detection system,” in *2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, 2018, pp. 1–5.

[27] M. Todisco, H. Delgado, and N. Evans, “A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients,” in *Odyssey 2016 - The Speaker and Language Recognition Workshop*, 06 2016.

[28] Z. Xie, W. Zhang, Z. Chen, and X. Xu, “A comparison of features for replay attack detection,” *Journal of Physics: Conference Series*, vol. 1229, p. 012079, 05 2019.