The Speaker and Language Recognition Workshop (Odyssey 2024)
18-21 June 2024, Quebec City, Canada

Spoofing detection in the wild: an investigation of approaches to improve generalisation

Anh-Tuan DAO, Nicholas Evans, Driss Matrouf

Laboratoire Informatique d'Avignon, Avignon Universite, France
EURECOM, Sophia Antipolis, France
{anh-tuan.dao, driss.matrouf}@univ-avignon.fr, evans@eurecom.fr

Abstract

The generalisation of spoofing detection solutions to spoofing attacks or recording conditions not seen in training data has been a focus since the inception of research in this area. We report our investigation of three strategies to improve upon generalisation, namely data augmentation, the fine-tuning of a pre-trained model, and a Siamese model with a cross-attention mechanism. When evaluated under domain-mismatched conditions, we show that these techniques are all effective in reducing model overfitting and in encouraging the learning of more generalisable models by capturing the (di)similarity between bonafide or spoofed test and known-to-be bonafide reference utterances. Evaluations using the *in-the-wild* dataset show that our model achieves a relative improvement of almost 60% compared to the best results reported in the literature.

Keywords: spoofing detection, generalisation learning, Siamese network

1. Introduction

Automatic speaker verification (ASV) technology offers a reliable and convenient means to biometric person recognition based on distinctive voice characteristics [1]. However, the rapid development of spoofing attack algorithms, which aim to deceive ASV systems with the impersonation of bonafide users, has raised concerns about security. Spoofing detection systems, also referred to as countermeasures (CMs) or presentation attack detectors, are nowadays deployed to combat attacks and to safeguard the integrity of ASV systems.

The ASVspoof [2] community, which includes broad participation from all over the world, has made significant advances in this field. Recent studies [3, 4] proposed the use of Graph Attention Networks to leverage the discriminative information in both temporal and spectral domains of a speech signal in order to discriminate spoofed from bonafide speech. Despite achieving impressive performance for the ASVspoof 2019 LA dataset, various studies have observed significant generalisation issues when spoofing detection systems are evaluated using datasets that are different to those with which they are trained. Müller et al. [5] demonstrated a lack of generalisation when spoofing detection models trained using the ASVspoof 2019 logical access (LA) databases are evaluated using the *in-the-wild* dataset. An equal error rate (EER) for a RawGAT-ST model rose from 1.2% for the evaluation partition of the same ASVspoof 2019 LA dataset to 37.1% when evaluated using the *in-the-wild* dataset.

A similar lack of generalisation was also reported for other models. Wang et al. [6] highlighted the influence of domain mismatch upon spoofing detection performance. They showed that AASIST-based models trained using the ASVspoof 2019 LA dataset erroneously classify a substantial proportion of the VoxCeleb2 dataset (of bonafide recordings) as spoofed.

Recent studies have employed fine-tuning strategies to mitigate model overfitting in spoofing detection. Tak et al. [7] fine-tuned the wav2vec 2.0 pre-trained model, a large-scale model for cross-lingual speech representation learning [8], using the ASVspoof 2019 LA training set. Experiments performed using the ASVspoof 2021 LA and DF datasets demonstrate that the use of pre-trained models leads to consistent improvements in generalisation to unseen spoofing attacks. Xie et al. [9] proposed a two-phase fine-tuning strategy using the wav2vec 2.0 pre-trained model and a Siamese network architecture. Their results show that the fine-tuning of a pre-trained model provides for powerful representation learning that improves spoofing detection performance for the ASVspoof 2019 LA dataset. While most existing methods [3, 4, 10] approach spoofing detection as a binary classification between bonafide and spoofed inputs, we assume that these approaches could lead to limitations in generalizing to unseen spoofing attacks. This is because classification models work best when the distribution of training and test data is similar. However, due to the rapid development of spoofing attack algorithms, it is challenging for the collection of training sets to keep pace with the development of new attack methods. Having recognised this, we have explored a blend of different techniques which improve generalisability to unseen spoofing attacks.

In this paper, we propose a novel approach to address the generalisation challenge. First, we establish new training and evaluation protocols using a combination of the ASVspoof 2019 LA dataset and the *in-the-wild* dataset. Second, we demonstrate the merit of a comprehensive data augmentation approach which is combined with the fine-tuning of pre-trained models. The augmentation strategies introduce additional variation to the training data so that models are less prone to overfitting and can learn more generalisable characteristics, which in turn improves detection reliability in the domain-mismatched scenario. Fine-tuning leverages the knowledge embedded in a pre-trained model, providing a solid foundation for the specialized spoofing detection task.

Third, we formulate spoofing detection as a similarity learning task [11]. We propose a Siamese-based cross-attention network which operates upon both the audio recording being analyzed and a bonafide reference recording. The model learns to map the two inputs to a single output which serves as an indication of similarity to the bonafide reference utterance - a prediction of whether the test utterance is bonafide or spoofed. This inherent characteristic enables Siamese networks to better generalize to new, previously unseen attacks. In addition, Siamese

10.21437/odyssey.2024-21

networks can effectively utilize the information from two inputs to make more informed decisions. The hypothesis is that use of a bonafide reference utterance is beneficial to the learning of characteristics which help distinguish bonafide test utterances from spoofs. The cross-attention mechanism [12] further enables the model to focus on the most relevant and discriminative information between test and bonafide reference utterances. Our Siamese model is trained according to two different strategies: (1) the bonafide reference utterance belongs to the same speaker as the training utterance (henceforth denoted Siamese-S); (2) the bonafide reference utterance can be from any speaker (denoted Siamese-A). While use of the same-speaker reference should results in the Siamese-S model better learning speaker characteristics, the Siamese-A model should better learn how spoofed utterances differ from bonafide utterances.

The remainder of this paper is organized as follows. Section 2 describes previous, related work. The proposed Siamese-based cross-attention model is described in Section 3. Experiments and results are presented in Sections 4-5. Finally, conclusions are presented in Section 6.

2. Related Works

In real-world environments, recording conditions can vary considerably. So too can the techniques used to implement spoofing attacks. Systems that are unable to generalize risk failing in these settings, potentially leading to security lapses. Addressing generalisation is crucial for the building of robust and trustworthy voice biometric systems which effectively combat the threat from increasingly sophisticated spoofing attacks. Although the ASVspoof 2019 LA dataset facilitates the evaluation of generalisation to unseen attacks, with an evaluation set generated using different spoofing algorithms to those used to generate training and development data, the evaluation of generalisation to varying recording conditions may require further exploration. Müller et al. [13] observed a correlation between the characteristics of non-speech segments and prediction labels. One interpretation of this finding might suggest vulnerabilities to unseen recording environments. There is hence a risk of overfitting to the acoustic conditions of data used for model training, instead of more generalisable characteristics which differentiate bonafide from spoofed speech.

The use of one-class classifiers is also a natural approach to improve generalisation. These approaches model only the bonafide class, for which training data is abundant, with sufficiently dissimilar utterances then being classified as spoofs. While the performance of one such approach reported in [14] is competitive, evaluation was performed using only the ASVspoof 2019 LA dataset, i.e. still an in-domain scenario. Other results reported in the literature show that the use of spoofed data for model training is still beneficial and, when evaluation is restricted to in-domain scenarios, the best two-class classifiers tend to outperform one-class competitors.

We adopt a relaxed training data policy for the ASVspoof 2019 LA dataset which prioritizes the learning of characteristics which generalize well to unseen scenarios. Our experiments demonstrate the effectiveness of data augmentation in promoting generalisation. We also investigate the fine-tuning of pre-trained models, which yields additional benefits. Finally, we introduce a Siamese-based cross-attention architecture which achieves, to the best of our knowledge, the best performance reported in the literature for the *in-the-wild* dataset.

3. Siamese-based cross-attention model

The Siamese architecture has been shown to be robust in handling unseen classes in verification systems [11]. Recent works [9, 15, 16] have adopted this architecture for speaker verification and spoofing detection, achieving promising results.

Our proposed Siamese model consists of two identical networks which share weights and process inputs from both the test utterance and a bonafide reference utterance. It then learns to map these inputs to a single output, a measure of their similarity (or dissimilarity). This output, indicating the degree of similarity to the bonafide reference, is used as a prediction of whether the test utterance is bonafide or spoofed. Figure 1 depicts an overview of the model architecture. Features are 80-dimensional Mel filterbank log-energy coefficients extracted from both utterances using frames of 400 ms with a 160 ms step size, and using a 512-tap FFT within the 20-7600 Hz frequency range. Subsequently, the mean is subtracted along the time axis.

We utilize the well-established ResNet architecture [17] which is known for its efficiency in image classification and recently-demonstrated success in speaker recognition [18]. The ResNet-generator structure in Figure 1 incorporates one convolution layer and four ResNet blocks, adhering to the ResNet34 architecture. The ResNet-generator output is flattened to attain a feature map H which is subsequently fed to the pair of cross-attention blocks. The specification of the ResNet34 architecture is presented in Table 1.

Given the feature maps H_{enroll} and H_{test} corresponding to the reference and test utterances respectively, we use cross-attention to capture inter-dependencies between the two feature maps, thereby identifying important features within each. Attention mechanisms have revolutionized Natural Language Processing (NLP) by enabling models to learn global dependencies between word embeddings in textual sequences. In our case, we treat feature maps as sequences of frequency bins, with cross-attention acting to emphasize those which are most informative for spoofing detection.

We used two cross-attention blocks to emphasize the most important features in H_{enroll} and H_{test}. The first block takes H_{enroll} as H_{source} and H_{test} as H_{support}, as shown in the Figure 2. The second block takes H_{test} as H_{source} and H_{enroll} as H_{support}. Each cross-attention block computes three matrices - a query matrix Q, a key matrix K, and a value matrix V - all the result of matrix multiplication upon H_{support} and H_{source}, as shown in Figure 2. Once these matrices are computed, attention scores are calculated. Scores reflect the relative importance of the information within each frequency bin. The scores are determined by multiplying the query matrix Q with the transposed key matrix K^T, and by scaling by d_k (dimensions of the query matrix Q). A softmax function is then applied to normalize the scores, ensuring they sum to 1. Finally, the attentive feature map H_{attn} is created by multiplying the attention scores with the value matrix V. The cross-attention layer output is computed according to

$$H_{\text{attn}} = \text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

The outputs of the cross-attention blocks H_{attn}, corresponding to the attentive feature maps of the reference and test utterances, are processed through pooling layers [19] and concatenated to obtain a combined embedding according to

$$h = \text{Concat}(\text{Pooling}(H_{attn\_enroll}), \text{Pooling}(H_{attn\_test})) \quad (2)$$

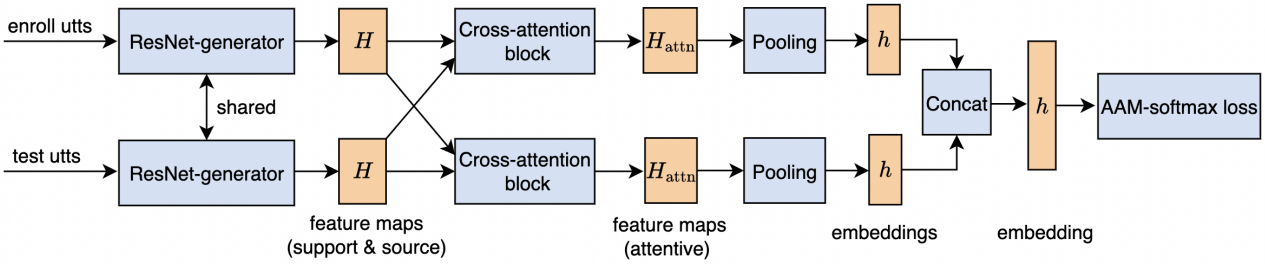

Figure 1: The architecture of our proposed Siamese-based cross-attention model. The output of the ResNet-generator, feature maps H, are fed into cross-attention blocks to calculate the attentive feature maps H_{attn}. Cross-attention blocks compute attentive feature maps H_{attn} for test and reference files, using each own feature map as H_{source}, as shown in the Figure 2. The colors in the figure indicate whether the block is a function or an embedding. Blue ones represent functions, while yellow ones represent feature maps or embeddings.

The combined embedding h is then fed into an AAM-softmax loss function [20] with *similarity* and *dissimilarity* classes.

4. Experimental Setup

4.1. Datasets and metrics

For all experiments reported in this paper, we used the ASVspoof 2019 LA database for training and development. All evaluation work was conducted using the *in-the-wild* dataset [5]. The use of an evaluation dataset different to that used for training and development helps to explore generalisation issues related to differences in spoofing attacks and acoustic conditions etc. For the fine-tuning of pre-trained models, we further used the VoxCeleb2 dataset [21], a popular resource for speaker recognition research. We provide a brief description of each dataset below.

- The **ASVspoof 2019 LA** dataset consists of three partitions: training, development, and evaluation. The training and development sets include bonafide in addition to spoofed utterances generated using 6 different spoofing algorithms (labeled A01-A06). The evaluation set also includes bonafide and spoofed utterances, with the latter being generated using 13 algorithms (A07-A19).
- The ***in-the-wild*** dataset contains approximately 31k audio recordings. Both bonafide and spoofed utterances are collected from English-speaking celebrities. Spoofed utterances are collected from 219 publicly available sources of deepfakes. Corresponding bonafide utterances for each celebrity are scraped from legitimate podcasts and recorded speeches available online.
- The **Voxceleb2** dataset is by far the most popular for speaker recognition research, boasting over 1 million utterances collected from 6,112 speakers. Its size facilitates the training of powerful models which can discriminate between target and non-target trial pairs in diverse and even challenging acoustic conditions.

Since we would otherwise have no use for the ASVspoof 2019 LA evaluation partition, and since it has the greatest diversity in terms of spoofing attack algorithms, which should be beneficial to genealisation, we use it for training. The remaining *training* and *development* partitions from ASVspoof 2019 LA were then pooled and used for validation. Though

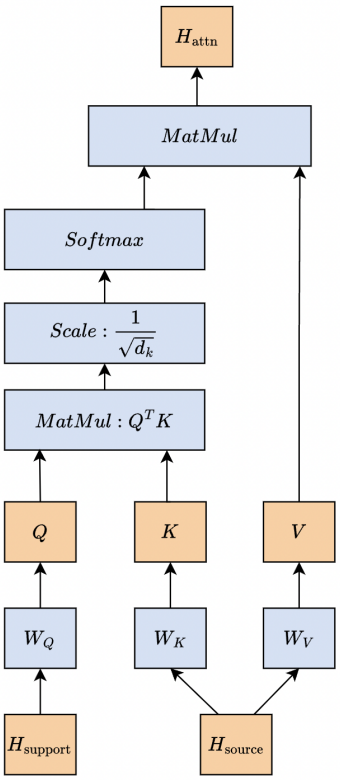

Figure 2: The cross-attention block computes the attentive feature map H_{attn} based on the feature maps of test and reference files. The attentive feature map of the test file uses the feature map of test file as H_{source}, the attentive feature map of the reference file uses the feature map of reference file as H_{source}.

Layer name	Output (C x F x T)	ResNet34
Conv2D	32 x 80 x T	$[3 \times 3, 32]$
ResBlock-1	32 x 80 x T	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$
ResBlock-2	64 x 40 x T/2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$
ResBlock-3	128 x 20 x T/4	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$
ResBlock-4	256 x 10 x T/8	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$
Flatten (C, F)	2560 x T/8	
Pooling	2560	Pooling
Dense	192	
AAM-Softmax	2	

Table 1: The architecture of ResNet model.

the ASVspoof 2019 LA dataset provides both ASV and CM protocols, here we are concerned only with spoofing detection. To align with the ASVspoof 2019 LA CM evaluation protocol where, for each test utterance there is a corresponding enrollment utterance for the same speaker, we constructed enrollment-test pairs for the *in-the-wild* dataset as well. *Enrollment* utterances (as in the sense of an ASV protocol) are used everywhere in this paper as bonafide *reference* utterances, for spoofing detection only. Details for each dataset are shown in Table 2. For all experiments, spoof detection performance is expressed in terms of the equal error rate (EER).

4.2. Data Augmentation

To tackle overfitting to the relatively clean acoustic conditions of the ASVspoof 2019 LA data, and to encourage generalisation to those of the *in-the-wild* database, we employed standard data augmentation techniques during training. We leveraged the MUSAN corpus and the real room impulse response (RIR) database [22, 23] to apply five diverse augmentation methods to each training utterance:

- Reverberation: utterances are convolved with real RIRs to simulate reverberation effects associated with propagation in various acoustic spaces.
- Speech: a summation of three to eight different-speaker utterances are added to each training utterance at signal-to-noise ratios (SNRs) of 13-20 dB.
- Music: randomly-selected music recordings from MUSAN are added to each training utterance at SNRs of 5-15 dB.
- Noise: randomly-selected noise recordings from MUSAN are added to each training utterance at SNRs of 0-15 dB.
- Spectral augmentation: SpecAugment [24] is applied to input log Mel-spectrograms, randomly masking between 0 and 10 frames in the time domain and between 0 and 8 frequency bin estimates.

4.3. Pre-training and fine-tuning

The fine-tuning of pre-trained models is a well-established technique to combat overfitting, especially when dealing with limited training data. In this section, we describe the pre-training and fine-tuning stages employed to effectively tackle the spoofing detection task.

Pre-training – We trained a ResNet model for speaker recognition using the VoxCeleb2 dataset. The model architecture is detailed in Table 1. The number of output units matches the number of speakers in the VoxCeleb2 dataset. Inputs are Mel-filterbank log-energy coefficients. This representation captures the spectral characteristics of each utterance. The input is then processed using a 2D convolution operation, which learns a feature map incorporating three dimensions: the number of channels (C); the number of frequency bins (F); the number of time frames (T). The feature map is subsequently passed through four consecutive ResNet blocks. These blocks extract progressively higher-level features from the data with skip connections. Each block includes convolution layers, followed by Batch Normalization and ReLU activation. Finally, a statistics pooling layer is used to process the learned feature map and to generate an embedding. This embedding represents the speaker's unique characteristics in a condensed form. The embedding is then fed into dense layers to produce the final output. An AAM-softmax function is used for the optimization of learnable model parameters.

Fine-tuning – We fine-tuned the pre-trained ResNet model using the ASVspoof 2019 LA evaluation set. Notably, the Siamese model shares the same architecture of the *ResNet-generator*, allowing us to directly load and fine-tune the model weights.

4.4. Implementation details

While training utterances are randomly segmented into single 4-second clips before being fed to the models, test utterances are evaluated using the *first* 4-second clip. The AAM-softmax loss function is used with hyperparameters *loss-scale* set to 30 and *loss-margin* set to 0.2. Our model is implemented using the PyTorch framework. For optimization, we employ the Adam optimizer with an initial learning rate of 0.0001 which decays by 3% per epoch. Fine-tuning is performed with a lower learning rate of 0.00001 to reduce model over-fitting. Training is conducted in batches of 32 samples. The model is trained for 80 epochs using 4 RTX 3090 GPUs. Each experiment was run three times with random seeds. Reported results are averages computed from the three runs.

5. Results

We report results for three sets of experiments. They are designed to assess the benefit to generalisation of: data augmentation; the fine-tuning of pre-trained models; the proposed Siamese model.

To examine the impact of data augmentation, we evaluated three methods: ResNet; AASIST [4]; RawGAT-ST [3]. We used the authors' implementations of AASIST and RawGAT-ST, and adapted them to our protocol and to incorporate data augmentation. Results for each model with and without data

Partition	Description	# Speakers	# Bonafide Utts	# Spoofing Utts
Train	ASVspoof 2019 LA eval	48	5370	63882
Dev	ASVspoof 2019 LA train + dev	30	4064	45096
Eval	*in-the-wild* dataset	58	19963	11816

Table 2: Summary of the dataset used in our experiments.

Systems	Data Aug	EER (%)
ResNet	No	37.90
	Yes	22.42
AASIST	No	46.07
	Yes	33.60
RawGAT-ST	No	43.03
	Yes	29.53

Table 3: Performance of different models in equal error rate (EER) evaluated on the *in-the-wild* dataset with and without data augmentation.

Systems	Fine-tuning	EER (%)
ResNet	No	22.42
	Yes	12.76
Siamese-S	No	20.90
	Yes	14.78
Siamese-A	No	18.50
	Yes	12.44

Table 4: Performance of different models in equal error rate (EER) evaluated on the *in-the-wild* dataset with and without fine-tuning. We conduct two training strategies for our Siamese models: (1) the bonafide reference utterance belongs to the same speaker as the training utterance (Siamese-S); (2) the bonafide reference utterance can be from any speaker (Siamese-A). Details of fine-tuning can be found in Section 4.3. All systems were performed with data augmentation.

augmentation are presented in Table 3. All methods performed poorly for the *in-the-wild* dataset without data augmentation. This finding corroborates those of Müller et al. [5]. Data augmentation is shown to be beneficial to generalisation, reducing error rates for the AASIST and RawGAT-ST models by 27% and 31.37% relative. Error rates for the ResNet model are lowest and, with an EER of 22.42%, the lowest overall when used with data augmentation. This represents a substantial relative improvement of 40.84% compared to performance without data augmentation. This could be because the larger AASIST and RawGAT-ST models are more prone to overfitting than the less complex RedNet model. Compared to the best performing RawGAT-ST model for the *in-the-wild* dataset as reported in [5], our ResNet model with data augmentation achieves a relative gain of 39.65% (22.42% vs. 37.15%).

Next, we assessed the benefit of fine-tuning, using ResNet and Siamese models. Given the benefit to generalisation, we trained both models using data augmentation. Results are presented in Table 4 and show additional improvements to generalisation. Fine-tuning is used to adapt the pre-trained model weights (pre-trained for speaker recognition) to the specific task (spoofing detection) and leads to better representation learning and generalisation. Relative improvements of 43% and 32.75% are obtained for the ResNet and Siamese-A models, respectively.

Also shown in Table 4 is a performance comparison for Siamese and ResNet models with data augmentation. The Siamese-A model shows better performance than the Siamese-S model under like-for-like conditions. These results indicate that the use of same-speaker references is detrimental to performance, with better results being obtained for the Siamese-A configuration. Without fine-tuning, the Siamese-A model exhibits better performance than the ResNet model, achieving a relative gain of 17.48%. After fine-tuning, performance for the Siamese-A model (12.44%) is still slightly better than that of the ResNet model (12.76%).

6. Conclusions

In this paper we report our investigation of three different approaches to improve generalisation in spoofing detection, namely data augmentation, the fine-tuning of pre-trained models, and a Siamese-based cross-attention model. Our results demonstrate notable improvements stemming from the use of all three techniques. Our best model, which uses the combination of all three, achieves what is, to the best of our knowledge, the lowest error rates reported in the literature for the *in-the-wild* dataset. This corresponds to a relative performance improvement of up to 60% in terms of equal error rate compared to the AASIST and RawGAT-ST models, state-of-the-art competitors according to results obtained for in-domain evaluation scenarios. Future work should couple the merits of cutting-edge AASIST and RawGAT-ST models for feature extraction with those of Siamese-based cross-attention models.

7. References

[1] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015.

[2] 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," *CoRR*, vol. abs/2109.00537, 2021.

[3] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu R. Kamble, Massimiliano Todisco, and Nicholas Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *ASVspoof workshop*, 2021.

[4] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. 2022, pp. 6367–6371, IEEE.

[5] Nicolas M. Müller, Pavel Czempin, Franziska Dieck-

mann, Adam Froghyar, and Konstantin Böttinger, "Does audio deepfake detection generalize?," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. 2022, pp. 2783–2787, ISCA.

[6] Xingming Wang, Xiaoyi Qin, Yikang Wang, Yunfei Xu, and Ming Li, "The DKU-OPPO system for the 2022 spoofing-aware speaker verification challenge," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. 2022, pp. 4396–4400, ISCA.

[7] Tak Hemlata, Todisco Massimiliano, Wang Xin, Jung Jeeweon, Yamagishi Junichi, and Evans Nicholas, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[8] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. 2022, pp. 2278–2282, ISCA.

[9] Yang Xie, Zhenchuan Zhang, and Yingchun Yang, "Siamese network with wav2vec feature for spoofing speech detection," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. 2021, pp. 4269–4273, ISCA.

[10] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher, "End-to-end anti-spoofing with rawnet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. 2021, pp. 6369–6373, IEEE.

[11] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, 1993, pp. 737–744.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[13] Nicolas M. Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, and Konstantin Böttinger, "Speech is silver, silence is golden: What do asvspoof-trained models really learn?," *CoRR*, vol. abs/2106.12914, 2021.

[14] You Zhang, Fei Jiang, and Zhiyao Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021.

[15] Yichi Zhang, Meng Yu, Na Li, Chengzhu Yu, Jia Cui, and Dong Yu, "Seq2seq attentional siamese neural networks for text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. 2019, pp. 6131–6135, IEEE.

[16] Umair Khan and Javier Hernando, "Unsupervised training of siamese networks for speaker verification," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*. 2020, pp. 3002–3006, ISCA.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016, pp. 770–778, IEEE Computer Society.

[18] Daniel Garcia-Romero, Gregory Sell, and Alan McCree, "Magneto: X-vector magnitude estimation network plus offset for improved speaker recognition," in *Odyssey 2020: The Speaker and Language Recognition Workshop, 1-5 November 2020, Tokyo, Japan*. 2020, pp. 1–8, ISCA.

[19] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. 2017, pp. 999–1003, ISCA.

[20] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, 2022.

[21] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. 2017, pp. 2616–2620, ISCA.

[22] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[23] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. 2017, pp. 5220–5224, IEEE.

[24] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*. 2019, pp. 2613–2617, ISCA.