INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Spoofing Attacker Also Benefits from Self-Supervised Pretrained Model

Aoi Ito[1]*, *Shota Horiguchi*[2]*

[1] Hosei University, [2] Hitachi, Ltd.

aoi.ito.8q@stu.hosei.ac.jp, shota.horiguchi.wk@hitachi.com

Abstract

Large-scale pretrained models using self-supervised learning have reportedly improved the performance of speech anti-spoofing. However, the attacker side may also make use of such models. Also, since it is very expensive to train such models from scratch, pretrained models on the Internet are often used, but the attacker and defender may possibly use the same pretrained model. This paper investigates whether the improvement in anti-spoofing with pretrained models holds under the condition that the models are available to attackers. As the attacker, we train a model that enhances spoofed utterances so that the speaker embedding extractor based on the pretrained models cannot distinguish between bona fide and spoofed utterances. Experimental results show that the gains the anti-spoofing models obtained by using the pretrained models almost disappear if the attacker also makes use of the pretrained models.

Index Terms: automatic speaker verification, anti-spoofing, self-supervised learning, wav2vec 2.0, HuBERT, WavLM

1. Introduction

Automatic speaker verification (ASV) is becoming a possible choice for secure authentication with the recent progress in its performance. However, ASV systems are exposed to the menaces of malicious attacks using speech synthesis, voice conversion, and replaying of recorded voice. To protect systems from such attacks, anti-spoofing methods are widely studied along with the growth of the community led by the ASVspoof challenges [1]. Although the boundaries are vague due to end-to-end modeling [2], anti-spoofing models generally consist of a feature extraction part and a classification part. As the input feature, hand-crafted features such as linear frequency cepstral coefficients [3] and constant Q cepstral coefficients [4] or features from neural networks such as SENet [5], DenseNet [6], and RawNet2 [2] are used in the literature.

The quality of features extracted from audio using neural networks is rapidly advancing with the self-supervised learning (SSL) paradigm; a lot of models have been proposed in the last few years such as wav2vec 2.0 [7], HuBERT [8], and WavLM [9]. They have shown greatly improved performance on various speech-related tasks such as speech recognition [10], speech enhancement [11], speaker identification [12], and emotion recognition [13]. Likewise, it has been reported that ASV anti-spoofing can also benefit from SSL models [14, 15, 16].

Although SSL models are powerful, training them from scratch consumes a lot of computational resources. For example, wav2vec 2.0, HuBERT, and WavLM have reportedly been trained using 64, 32, and 16 NVIDIA V100 GPUs, respectively, even for the smallest BASE model of each. These computing environments are not necessarily on a scale that is readily available to everyone. Therefore, it is a common practice to finetune publicly available pretrained models on the Internet when using them for one's own applications.

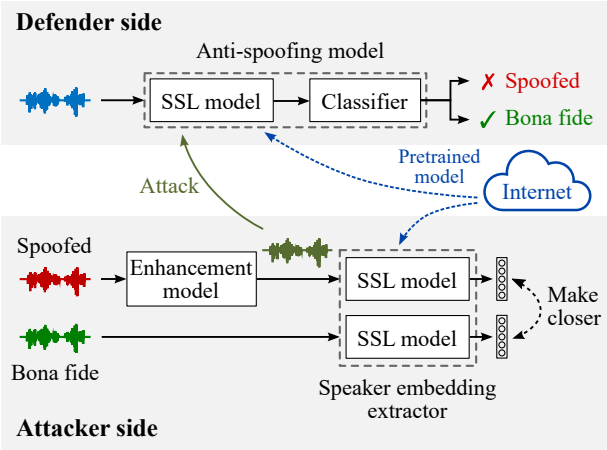


Figure 1: *Overview of evaluation framework*

Here, although the accuracy of ASV anti-spoofing has indeed been reportedly improved by SSL models, it is easy to imagine that the attacker side can also take advantage of the power of these models. There is no previous investigation of the extent to which anti-spoofing loses the gains obtained by SSL models when attackers also use them. Also, if an ASV anti-spoofing system is developed using a publicly available pretrained model, an attacker can also access the same model. The security risks from the same pretrained models being used in attacks against anti-spoofing systems have also not been explored before.

This paper aims to provide answers to two questions: i) if an attacker leverages SSL models, will they maintain the anti-spoofing performance reportedly improved by them, and ii) is there any disadvantage to anti-spoofing if the attacker uses the same pretrained model on the Internet as the defender even when it is finetuned? For this purpose, we propose a method for training the attacker, as illustrated in Figure 1. In the method, a speaker embedding extractor based on SSL models is first trained. Then, an enhancement model that improves the deception ability of spoofed utterances is trained to make the speaker embeddings extracted from bona fide and *enhanced* spoofed utterances closer. The proposed method is evaluated using various combinations of attacker-defender pairs on three ASVspoof challenge datasets.

*The authors equally contributed to this work. This work has been done during Aoi Ito's internship at Hitachi, Ltd.

10.21437/Interspeech.2023-270

2. Related work

If the attacker has prior knowledge of an anti-spoofing model, adversarial attacks can be attempted: white-box attacks when the parameters of the model are known [17] and black-box attacks when only input-output pairs of the model are known [18] (please refer to [19] for a literature review). The performance of ASV with SSL models was also degraded by adversarial attacks [20], but the models were not finetuned. When a model is fine-tuned, it is not possible to directly calculate such adversarial attacks. One possible attack in such a case is when a pretrained model has a backdoor in it due to weight poisoning [21], but this is not considered in this paper because we assume the commonly used SSL model. Another possibility is to utilize the transferability of adversarial attacks; successful attacks on a certain model are also likely to fool other models [22]. While attacks by adversarial samples are an important issue for any machine learning model, the scope of this paper is to verify whether the attacker can also benefit from the power of SSL models, so generating such samples artificially is out of scope for this paper.

The paradigm used in this paper is highly related to verification-to-synthesis (V2S) [23], in which a voice conversion model is trained to fool a pretrained speaker classification model. There are three differences from V2S: i) we enhance utterances that are already made to spoof target speakers, ii) we do not apply any regularization to enhanced utterances to preserve their phonetic properties, and iii) we directly compare a pair of speaker embeddings instead of the output from the classifier and the desired speaker label, which makes the method applicable to any speaker.

3. Method

We assume the situation depicted in Figure 1. The defender side builds an anti-spoofing model by using an SSL model on the Internet as a frontend. The anti-spoofing model performs a two-way classification that distinguishes whether an input is a bona fide or spoofed utterance. The attacker side aims to obtain a spoofing model that transforms an already spoofed piece of audio by, for example, voice conversion or text-to-speech synthesis, to enhance its spoofing ability. With a speaker embedding extractor based on the SSL model, the spoofing model is trained so that the embedding extracted from a spoofed input is indistinguishable from that from a bona fide input. We will describe the implementation of each side in detail in the following subsections.

3.1. Defender side

For the anti-spoofing model, we used a simplified version of the RawNet2-based architecture [16], in which the SincNet frontend [24] was replaced by an SSL model. Table 1 shows the detailed configuration of the model. Given 64,600 samples of audio ($\sim$4 seconds), the SSL frontend first extracts 201-length 768-dimensional speech representations. The following dimensionality reduction layer and six-stacked residual blocks further convert the representations to perform two-way classification, *i.e.*, spoofed vs. bona fide, in the last layer. The model was trained using the standard cross entropy loss. Note that the parameters of the SSL frontend were initialized with those of the pretrained model and jointly optimized with the backend in an end-to-end manner.

Table 1: *Architecture of anti-spoofing model. Input to model is 64600-length waveform. Output shape corresponds to number of channels, frames, and frequency bins, respectively. BN and SeLU refer to batch normalization and scaled exponential linear unit, respectively.*

Layer	Output shape	Configuration
SSL frontend	(1, 201, 768)	wav2vec 2.0/HuBERT/WavLM
Dimensionality reduction	(1, 201, 128)	128-dim fully connected
	(1, 67, 42)	3×3 max pooling
	(1, 67, 42)	BN & SeLU
Residual block	(32, 67, 42)	$\begin{bmatrix} \text{3x3 conv, 32-ch} \\ \text{BN \& SeLU} \\ \text{3x3 conv, 32-ch} \end{bmatrix} \times 2$
Residual block	(64, 67, 42)	$\begin{bmatrix} \text{3x3 conv, 64-ch} \\ \text{BN \& SeLU} \\ \text{3x3 conv, 64-ch} \end{bmatrix} \times 4$
Classification	64	67×42 global average pooling
	2	2-dim fully connected

3.2. Attacker side

The attacker side attempts to transform spoofed recordings into ones that the SSL model cannot distinguish from bona fide recordings. The training of the spoofing model is two-staged. First, a speaker embedding extractor is constructed on the basis of the model pretrained using SSL, and then an enhancement model to improve the input's spoofing ability is trained using the extractor. If the enhancement model can be trained to be able to fool the extractor, then an anti-spoofing model based on the same SSL model could be fooled as well.

The speaker embedding extractor f_{embed} was trained to classify utterances on the basis of their speaker IDs. Given an input utterance, the SSL frontend first computes frame-level embeddings, which are then aggregated by average pooling along the time axis to obtain an utterance-level embedding. Following the conventional study [25], the entire network was optimized using additive angular margin (AAM) softmax loss [26].

With the well-trained speaker embedding extractor, the enhancement model f_{enh} was trained to convert spoofed utterances to ones whose speaker embeddings are not distinguishable from those extracted from bona fide utterances. During training, a pair of bona fide and spoofed utterances is used to train the model. A spoofed utterance $\boldsymbol{x}_{\text{spoof}}$ is first fed to the spoofing model to convert it to an enhanced one:

$$\boldsymbol{x}_{\text{enh}} = f_{\text{enh}}\left(\boldsymbol{x}_{\text{spoof}}\right). \tag{1}$$

For the enhancement model f_{enh}, we used Conv-TasNet [27] to convert the input audio in the time domain. Then, speaker embeddings are extracted from each bona fide and *enhanced* spoofed utterances. The network is optimized to minimize the angle between those embeddings by using the following loss:

$$\mathcal{L} = 1 - \cos\left(f_{\text{embed}}\left(\boldsymbol{x}_{\text{enh}}\right), f_{\text{embed}}\left(\boldsymbol{x}_{\text{bonafide}}\right)\right), \tag{2}$$

where $\boldsymbol{x}_{\text{bonafide}}$ is a bona fide utterance of the speakers who the spoofed utterance $\boldsymbol{x}_{\text{spoof}}$ is pretending to be and $\cos(\cdot,\cdot)$ denotes the cosine similarity between two arguments. Note that the parameters of the speaker embedding extractor were frozen during the training of the enhancement model; otherwise, it will fall into a trivial solution that, for example, always outputs the same embedding regardless of the input.

During evaluations, *enhanced* spoofed utterances obtained using (1) are fed to the anti-spoofing models described in Section 3.1 instead of the original spoofed utterances.

4. Experimental settings

4.1. Dataset

We consider two scenarios in this paper. The first scenario is that an attacker and a defender use a different dataset for the training of each side's model. To meet this purpose, each of the training and development sets of the ASVspoof 2019 logical access (LA) database [28, 29] was divided into two portions to train spoofing models and anti-spoofing models, respectively. For the bona fide utterances, we assigned half of each speaker's utterances to the attacker and the other half to the defender. For the spoofed utterances, assuming that an attacker does not have prior knowledge of the specific method that a defender is taking into consideration, we divided them based on their systems: A01, A03, A05 for the attacker and A02, A04, A06 for the defender[1]. The second scenario is that an attacker has access to some of the defender's data, e.g., the defender makes use of publicly available datasets. In this scenario, the defender uses the whole ASVspoof 2019 LA database, while the attacker uses the same portion as the first scenario.

For the evaluation, we used the test set of the ASVspoof 2019 LA database as a clean dataset, in which spoofing attacks are based on speech synthesis and voice conversion. We also used the ASVspoof 2021 LA and deepfake (DF) databases [30] for more noisy and realistic trials. The 2021 LA database is based on the same attack algorithms as the 2019 LA database, but the effects of encoding and transmission over the telephone are also taken into account. The 2021 DF database focuses on the distortion caused by compressing and restoring audio through various codecs. Although there is some discussion about the appropriateness of the original ASVspoof datasets [31], we used them as they are. The models' performance was evaluated using an equal error rate (EER).

4.2. Model configuration

As the SSL models, we used wav2vec 2.0 BASE [7], HuBERT BASE [8], WavLM BASE, and WavLM BASE+ [9] models. For simplicity, the word "BASE" will be omitted hereafter. The `TorchAudio` [32] implementations of wav2vec 2.0 and HuBERT and the official implementation of WavLMs[2] were used in our experiments. Each model has approximately 95 million parameters. The wav2vec2.0, HuBERT, and WavLM were trained with the concatenation of *train-clean-100*, *train-clean-360*, and *train-other-500* from the LibriSpeech dataset [33], and WavLM+ was trained with the Libri-Light [34], GigaSpeech [35], and VoxPopuli [36] datasets.

The anti-spoofing models were trained on the four types of SSL models above. They were trained using the Adam optimizer [37] with a fixed learning rate of 1×10^{-6} for at most 100 epochs. The batch size was set to 32. As the baseline without an SSL frontend, we also used the official RawNet2 recipe from the ASVspoof 2021 baseline[3]. No data augmentation techniques were applied during training. The inputs to each model were aligned to 64,600 samples by cropping and/or repeating the original utterances.

[1]This split was to avoid giving unfair advantages to the attacker since the spoofing systems used for A04 and A06 were also used in the test set.

[2]`https://github.com/microsoft/unilm/tree/master/wavlm`

[3]`https://github.com/asvspoof-challenge/2021`

Table 2: *Speaker verification results on VoxCeleb1 dataset*

	EER (%)		
SSL model	VoxCeleb1	VoxCeleb1-E	VoxCeleb1-H
wav2vec 2.0	2.49	2.91	6.05
HuBERT	2.85	3.15	6.49

For speaker embedding extractors, we simply finetuned the wav2vec 2.0 and HuBERT models using VoxCeleb2 [38] and evaluated them using VoxCeleb1. No extra embedding layer as the backend was introduced; thus, the dimensionality of the speaker embedding was 768. Each model was optimized to minimize the AAM softmax loss with a margin of 0.3 and a scale of 15 using the Adam optimizer for 6 epochs ($\sim$100k iterations). The learning rate was linearly increased from zero to 1×10^{-5} for the first $10\,\%$ of iterations, kept unchanged for the next $40\,\%$ of iterations, and then linearly decayed to zero for the final $50\,\%$ of iterations. Here also, we did not apply any data augmentation techniques during training.

The spoofing models based on Conv-TasNet were trained using each speaker embedding extractor as a backend. We used a Conv-TasNet model that consists of three repetitions of eight-stacked convolutional blocks with different dilations, which was the best configuration in the original paper [27]. The training was conducted for 300 epochs using the Adam optimizer with a fixed learning rate of 1×10^{-5} and a batch size of 8 without data augmentation.

5. Results

5.1. Preliminary results on attacker side

Before discussing the main results, we report the performance of the model trained for the attacker side. Table 2 shows the EERs of the speaker embedding extractors on the VoxCeleb1 dataset. We use three cleaned splits of the dataset: the original test set (VoxCeleb1 in Table 2), the extended test set (VoxCeleb1-E), and the hard test set (VoxCeleb-H). Although the quality of the speaker embeddings is not the main focus of this paper, we note that these values are comparable, though not the best, to previously reported values [25].

5.2. Main results

Table 3 shows the EERs for the cases when the anti-spoofing models were trained using the whole ASVspoof 2019 LA training set. Without spoofing enhancement models, the anti-spoofing models with the SSL frontend significantly outperformed the RawNet2 baseline on all the evaluation datasets, consistent with the previous study [16]. From the results on the 2019 LA test set (Table 3a), the absolute EERs of RawNet2-based anti-spoofing were significantly increased by the spoofing enhancement, while those of the SSL-based anti-spoofing were hardly affected. This suggests that SSL models are not fooled by spoofing enhancement in clean conditions where the training and evaluation data do not diverge. On the other hand, the results on the 2021 LA and DF test sets in Tables 3b and 3c show that not only RawNet2-based anti-spoofing but also SSL-based anti-spoofing were degraded by spoofing enhancement. The performance gains that the defender side obtained by using SSL models were almost completely lost when the attacker side also used the SSL models. Even so, since the RawNet2 base-

Table 3: *EERs (%) when anti-spoofing models were trained using whole ASVspoof 2019 LA training set.*

(a) *ASVspoof 2019 LA test set*

	Spoofing enhancement		
Anti-spoofing	None	wav2vec 2.0	HuBERT
RawNet2	17.49	60.60	57.03
wav2vec 2.0	0.81	0.60	0.71
HuBERT	1.62	2.86	2.83
WavLM	1.03	2.68	2.53
WavLM+	0.44	0.24	0.23

(b) *ASVspoof 2021 LA test set*

	Spoofing enhancement		
Anti-spoofing	None	wav2vec 2.0	HuBERT
RawNet2	18.06	67.68	63.70
wav2vec 2.0	7.20	16.57	16.47
HuBERT	4.89	25.94	23.95
WavLM	7.99	33.95	32.59
WavLM+	7.55	26.94	25.66

(c) *ASVspoof 2021 DF test set*

	Spoofing enhancement		
Anti-spoofing	None	wav2vec 2.0	HuBERT
RawNet2	24.01	68.46	65.77
wav2vec 2.0	10.31	25.49	26.36
HuBERT	18.93	46.22	45.49
WavLM	16.14	46.71	45.97
WavLM+	11.08	33.63	32.24

Table 4: *EERs (%) when anti-spoofing models were trained using portion of ASVspoof 2019 LA training set.*

(a) *ASVspoof 2019 LA test set*

	Spoofing enhancement		
Anti-spoofing	None	wav2vec 2.0	HuBERT
RawNet2	10.93	24.28	24.62
wav2vec 2.0	0.82	0.60	0.67
HuBERT	1.70	2.12	2.09
WavLM	0.94	3.05	2.95
WavLM+	0.73	0.30	0.28

(b) *ASVspoof 2021 LA test set*

	Spoofing enhancement		
Anti-spoofing	None	wav2vec 2.0	HuBERT
RawNet2	11.64	29.31	29.45
wav2vec 2.0	8.28	20.05	19.51
HuBERT	5.40	15.25	13.90
WavLM	21.20	38.06	38.00
WavLM+	10.76	29.20	28.10

(c) *ASVspoof 2021 DF test set*

	Spoofing enhancement		
Anti-spoofing	None	wav2vec 2.0	HuBERT
RawNet2	24.47	49.76	48.24
wav2vec 2.0	10.03	21.16	22.17
HuBERT	15.07	38.40	38.16
WavLM	14.78	42.30	42.48
WavLM+	10.56	31.95	30.49

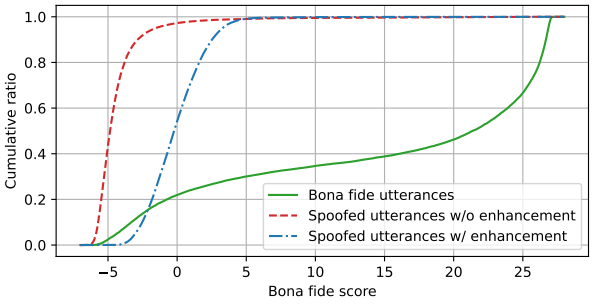
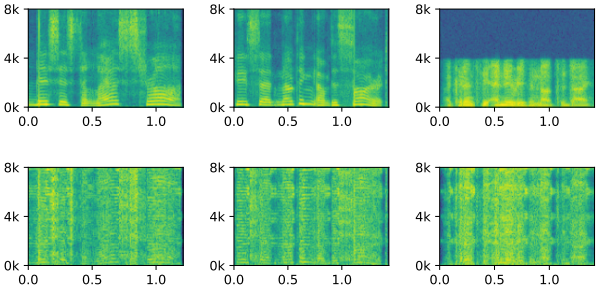


Figure 2: *Cumulative distributions of bona fide score on ASVspoof 2021 LA test set. Anti-spoofing model and enhancement model are both based on wav2vec 2.0.*



(a) *LA_E_8161161* $(-5.7 \rightarrow 25.5)$ (b) *LA_E_5770469* $(-5.6 \rightarrow 23.7)$ (c) *LA_E_4822201* $(-5.6 \rightarrow 21.1)$

Figure 3: *Examples of spoofed utterances in ASVspoof 2021 LA test set before (top) and after (bottom) enhancement. Horizontal and vertical axes correspond to time and frequency, respectively. Anti-spoofing and enhancement models are both based on wav2vec 2.0. Values in brackets show how bona fide score was changed by enhancement.*

line also suffers from performance degradation due to spoofing enhancement, the use of the SSL model for defense is recommended. From the results, no particular performance degradation was observed when using the same pretrained model for the attacker and the defender, e.g., a $16.57\,\%$ EER on the 2021 LA test set when both sides used wav2vec 2.0. This means that it is a possible option to use publicly available SSL models on the Internet for defense.

Table 4 shows the EERs for the case when the anti-spoofing models were trained using the portion of the 2019 LA training set. The trend in the results is the same as in Table 3; spoofing enhancement did not affect the results for the clean condition (2019 LA) that much but degraded those for the realistic conditions (2021 LA and DF).

Figure 2 shows the cumulative distribution of the bona fide score, which was obtained as the logit of the bona fide class, for the 2021 LA set. Both the anti-spoofing and spoofing enhancement models are based on wav2vec 2.0. It is clearly observed that the spoofing enhancement shifted the distribution of the spoofed utterances to the right, *i.e.*, increased the bona fide score.

Figure 3 shows examples of the original spoofed utterances and their conversions of which the bona fide scores were largely increased by the enhancement model based on wav2vec 2.0. While spoofing enhancement makes utterances fool an anti-spoofing model, their harmonic structures are rarely preserved. Spoofing enhancement that also preserves the naturalness of transformed utterances (like V2S [23]) is left to future work.

6. Conclusion

In this study, we investigated whether the performance improvement of an anti-spoofing model obtained by using an SSL model is real, under the condition that the attacker also has access to the SSL model. Experiments revealed that the attacker could also benefit from SSL models, thereby eliminating most of the benefits the defender gains from them. We also found no significant EER degradation for the attacker side from using the same pretraining model as the defender side, indicating that both sides should simply use a stronger SSL model. Future work will include a countermeasure for attackers in which pretrained SSL models are utilized.

7. References

[1] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[2] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. ICASSP*, 2021, pp. 6369–6373.

[3] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. BTAS*, 2013, pp. 1–8.

[4] 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.

[5] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-spoofing with squeeze-excitation and residual networks," in *Proc. INTERSPEECH*, 2019, pp. 1013–1017.

[6] C. Zhang, J. Cheng, Y. Gu, H. Wang, J. Ma, S. Wang, and J. Xiao, "Improving replay detection system with channel consistency densenext for the asvspoof 2019 challenge." in *Proc. INTERSPEECH*, 2020, pp. 4596–4600.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460, 2021.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[10] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," in *Proc. ICASSP*, 2019, pp. 7694–7698.

[11] Z. Huang, S. Watanabe, S.-w. Yang, P. García, and S. Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *Proc. ICASSP*, 2022, pp. 6837–6841.

[12] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. INTERSPEECH*, 2021, pp. 1509–1513.

[13] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. INTERSPEECH*, 2021, pp. 3400–3404.

[14] Y. Xie, Z. Zhang, and Y. Yang, "Siamese network with wav2vec feature for spoofing speech detection." in *Proc. INTERSPEECH*, 2021, pp. 4269–4273.

[15] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in *Proc. Odyssey*, 2022, pp. 100–106.

[16] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. Odyssey*, 2022, pp. 112–119.

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. ICLR*, 2014.

[18] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM ASIACCS*, 2017, pp. 506–519.

[19] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *Proc. INTERSPEECH*, 2020, pp. 4213–4217.

[20] H. Wu, B. Zheng, X. Li, X. Wu, H.-Y. Lee, and H. Meng, "Characterizing the adversarial vulnerability of speech self-supervised learning," in *Proc. ICASSP*, 2022, pp. 3164–3168.

[21] K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pretrained models," in *Proc. ACL*, 2020, pp. 2793–2806.

[22] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-box attacks on spoofing countermeasures using transferability of adversarial examples." in *Proc. INTERSPEECH*, 2020, pp. 4238–4242.

[23] T. Nakamura, Y. Saito, S. Takamichi, Y. Ijima, and H. Saruwatari, "V2S attack: building DNN-based voice conversion from automatic speaker verification," in *Proc. ISCA SSW*, 2019, pp. 161–165.

[24] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. SLT*, 2018, pp. 1021–1028.

[25] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *Proc. ICASSP*, 2022, pp. 7967–7971.

[26] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE TPAMI*, vol. 44, no. 10, pp. 5962–5979, 2022.

[27] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[28] 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," in *Proc. INTERSPEECH*, 2019, pp. 1008–1012.

[29] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[30] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof Challenge*, 2021, pp. 47–54.

[31] N. M. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is silver, silence is golden: What do ASVspoof-trained models really learn?" in *Proc. ASVspoof Challenge*, 2021.

[32] Y.-Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang *et al.*, "TorchAudio: Building blocks for audio and speech processing," in *Proc. ICASSP*, 2022, pp. 6982–6986.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[34] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-Light: A benchmark for ASR with limited or no supervision," in *Proc. ICASSP*, 2020, pp. 7669–7673.

[35] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," in *Proc. INTERSPEECH*, 2021, pp. 3670–3674.

[36] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. ACL-IJCNLP*, 2021, pp. 993—-1003.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[38] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.