Interspeech 2024
1-5 September 2024, Kos, Greece

# Spoofed Speech Detection with a Focus on Speaker Embedding

*Hoan My Tran*[1], *David Guennec*[1], *Philippe Martin*[1], *Aghilas Sini*[2], *Damien Lolive*[1], *Arnaud Delhay*[1], *Pierre-François Marteau*[3]

[1]Univ Rennes, IRISA, CNRS, France; [2]Université Le Mans, LIUM, France
[3]Université Bretagne Sud, IRISA, CNRS, France

{hoan.tran, david.guennec, philippe.martin, damien.lolive, arnaud.delhay, pierre-francois.marteau}@irisa.fr, aghilas.sini@univ-lemans.fr

## Abstract

Self-Supervised Learning (SSL) models excel as feature extractors in downstream speech tasks, including the increasingly crucial area of spoof speech detection due to the rise of audio deepfakes using Text-To-Speech (TTS) and Voice Conversion (VC) technologies. To address this issue, we propose a novel approach that relies on speaker embedding using a finetuned WavLM model with layer-wise attentive statistics pooling combined to a supervised contrastive learning and cross-entropy loss. Evaluation on Logical Access (LA) and DeepFake (DF) tasks on ASVspoof 2019 and 2021 highlights its potential in detecting audio deepfakes, with the contrastive loss producing more stable results among test sets.

**Index Terms**: anti-spoofing, self-supervised learning, supervised contrastive learning, speaker recognition.

## 1. Introduction

While speech synthesis, voice cloning and voice conversion offer many benefits, such as in the accessibility domain, improvements in perceptual quality over recent years raised concerns regarding misuses such as identity theft and fraud. The necessity of countermeasures for both speaker verification and generated audio (thus fake audio) detection is therefore becoming an important research topic. In particular, the ASVspoof community actively develops Presentation Attack Detection (PAD) systems and public datasets to address these challenges [1, 2, 3, 4].

Speaker verification leverages voice biometrics (pitch, formant frequencies, vocal resonances, speaking styles) for identification in areas like phone banking and access control [5]. Limitations like background noise sensitivity and spoofing attacks hinder performance.

Several countermeasures have been proposed to address spoofing detection in ASVspoof challenges, including baseline techniques such as those described in [6, 7]. Recent advancements in Graph Neural Networks (GNNs) models have shown promising results in improving spoofing detection performance, as demonstrated in [8, 9]. Additionally, [10] explores the use of vision transformers (ViT) [11] with supervised contrastive loss, while [12] investigates the application of data augmentation techniques like RawBoost [12] to enhance the robustness of spoofing detection solutions.

Advancements in SSL models for speech representations, like HuBERT [13], Wav2Vec 2.0 [14], and WavLM [15], have revolutionized various speech tasks, including speech recognition, speaker identification, and sentiment analysis [16]. Compared to traditional features (e.g., MFCC, LFCC, CQCC), these models extract richer information, leading to significant gains in spoof speech detection.

Several recent studies have demonstrated the effectiveness of Wav2vec 2.0 [17] as a front-end for spoofed speech detection. This approach leverages attention-based pooling layers [18], particularly Attentive Statistics Pooling (ASP) [19], to extract rich speaker embedding information. Notably, ASP has been shown to be particularly effective in this task [20]. Additionally, Guo et al. [21] explored the use of WavLM as a feature extractor and demonstrated its efficiency. Li et al. [22] employed WavLM as their discriminator feature encoder for detecting synthesized and spoofed speech. These models, pretrained on massive speech datasets, benefit from transfer learning, allowing them to learn general representations applicable to downstream tasks like spoofing detection. However, instead of full fine-tuning, which can be computationally expensive, an optimal configuration for the fine-tuning process is crucial. Layer-wise analysis, as described in [15, 23, 24, 25], reveals that various layers capture distinct acoustic and linguistic features, offering valuable insights into selecting the most informative layers for partial fine-tuning and highlighting the potential of the approach in improving spoofing detection performance.

Continuing in this trend, the present study investigates the utilization of acoustic information for speaker representation learning within the context of spoofed speech detection. We leverage the pre-trained WavLM model and strategically select specific transformer layers for fine-tuning. Several configurations are considered depending on the information provided by the layer. Furthermore, in order to enhance the model's ability to discriminate between genuine and spoofed speech representations, we combine supervised contrastive loss [26] with cross-entropy loss during the fine-tuning process [27]. The supervised contrastive loss guides the feature encoder model to learn distinct embeddings for bona fide and spoof speech, while the cross-entropy loss aids in the final classification task. Our new method, which relies on WavLM as a feature extractor, thus explores 3 main points: (1) strategic choice of the contextual transformer encoders to maximize information related to speaker embedding in our classifier, (2) fine-tuning only these WavLM layers alongside our classifier on spoofing data and (3) adding a supervised contrastive loss in order to help feature extractor distinguish better spoofed and bona fide features. Our method is then tested on data from both 2019 and 2021 editions of the ASVspoof Challenge. We confirm that selecting carefully the embedding layers improves results on the spoofing detection task and show that the contrastive loss provides more stable results among evaluation sets.

We start by presenting our model architecture in section 2. Our experimental setup is then presented in section 3. Finally, section 4 discusses our results before concluding in section 5.


10.21437/Interspeech.2024-481

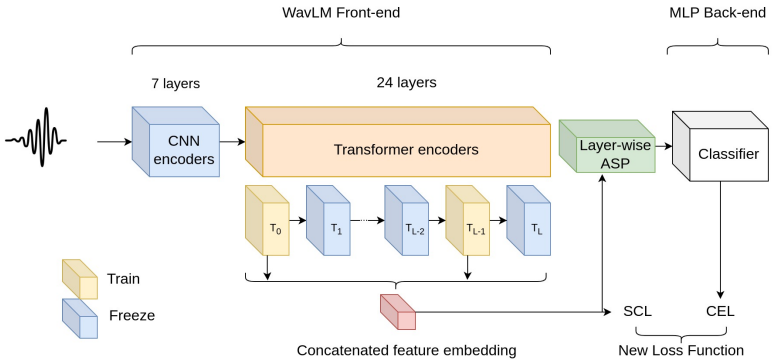


Figure 1: *The global model architecture is split into two main parts: the WavLM front-end is used to extract embeddings that are passed to the classifier. The embedding vectors extracted from the different layers are merged using ASP. Two losses are used to train the model: the Cross-Entropy Loss and the Supervised Contrastive Loss (SCL). The latter helps to improve the distinction between the two classes, namely bona fide and spoof speech samples.*

## 2. Model Architecture

The architecture investigated in this article is based around a 2-blocks pipeline as presented in figure 1: WavLM for feature extraction and a classifier for spoofed audio detection. First an audio is provided to the WavLM model which is composed of 7 CNN layers for encoding acoustic information. It is then passed to a stack of 24 transformer layers. Embeddings extracted by each individual layer are then merged using ASP and passed to the spoofed audio classifier. In all experiments presented in this paper, the CNN encoder is frozen. As detailed below, various transformer layers may also be frozen. The different components of the model are detailed hereafter.

### 2.1. Feature extraction with WavLM

Unlike other pre-training methods that may not prioritize speaker information, WavLM is specifically designed with an emphasis on preserving speaker identity. This is achieved through its architecture which is built upon the HuBERT framework with mechanisms to capture speaker information. The pre-training stage includes an "utterance and speaker contrastive loss" which encourages the model to distinguish between different speakers while understanding the spoken content. The denoising objective in WavLM's pre-training strengthens the model's ability to handle noise. This is crucial for speaker embedding tasks as real-world audio recordings often contain background noise or environmental variations. By being robust to noise, WavLM can extract speaker-specific features more reliably, leading to better speaker embeddings [28]. WavLM is trained on a massive dataset of speech recordings (94k hours) [29, 30, 31], which allows it to learn intricate details within the audio data. This includes capturing the subtle variations in voice characteristics that differentiate individual speakers. This extensive training provides a strong foundation for accurate speaker embedding.

### 2.2. Attentive Statistics Pooling

The extracted embeddings from WavLM transformer layers are concatenated and run through Attentive Statistical Pooling (ASP) which allows to extract higher-order features (standard deviation, average characteristics of the speaker) more suitable for speaker discrimination [19]. ASP takes advantage of both the attention mechanism and statistics pooling to produce weighted mean and standard deviation vectors. The attention mechanism identifies important frames, assigning higher weights to them, while statistics pooling captures the overall distribution and variability of frame-level features. This combined approach aims to enhance the discriminative power of utterance-level features for speaker recognition.

### 2.3. Spoofed audio Classifier

The ASP output features are passed through a binary classifier which distinguishes between spoofed and bona fide audio. The classifier is an MLP that takes the embeddings as input, with one hidden layer of 512 neurons in size (ReLU activation function) and given an output of size 2. Training is done using a standard binary Cross Entropy Loss function (CEL) which ensures that the learned representations are informative for the given task labels:

$$\mathcal{L}_{\text{CEL}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i) \qquad (1)$$

where $N$ is the total number of examples in the minibatch, $y_i \in \{0, 1\}$ is the ground truth label for sample $i$, $p_i$ is the predicted probability of sample $i$ belonging to class 0 (e.g. bona fide) and $(1 - p_i)$ the probability of class 1 (e.g. spoof).

### 2.4. Contrastive learning

In order to further discriminate between spoofed and bona fide audio, we add a contrastive loss to the cross-entropy classification loss to jointly optimize the representation learning (WavLM transformer layers that are not frozen) and the classification task (the spoofed audio classifier). As in [26], the Supervised Contrastive Loss (SCL) is used to encourage the learning of discriminative representations by embedding examples into a high-dimensional space. This means it makes used of both labeled data and the contrastive principle to train the model. In our case, it encourages the model to learn representations pushing away bona fide and spoofed samples from each other.

A temperature parameter $\tau$ is used to modulate the scale of similarities between examples, influencing the sharpness of the similarity distribution. Through the dot product of feature vectors $\Phi(x_i) \cdot \Phi(x_j)$, the similarity between examples $i$ and $j$ is measured. The difference is then scaled by the temperature

| Evaluation set | Bona fide | Spoof |
|---|---|---|
| LA 2019 | 7355 (10.33%) | 63882 (89.67%) |
| LA 2021 | 14816 (10.33%) | 133360 (89.67%) |
| DF 2021 | 14869 (2.78%) | 519059 (97.22%) |

Table 1: *Number of audio in the 2019 and 2021 datasets used for the ASVSpoof Challenge.*

parameter, thus influencing the degree of discrimination in the learned representation:

$$z(i,j) = \frac{\Phi(x_i) \cdot \Phi(x_j)}{\tau} \tag{2}$$

By employing an indicator $\mathbb{1}_{y_i=y_j}$, only positive pairs (examples from the same class) contribute to the loss, promoting the aggregation of similar examples and the separation of dissimilar ones.

The loss is computed by negating and summing the logarithm of the ratio of similarities between positive and negative pairs, ensuring that similar examples are brought closer together while dissimilar ones are pushed further apart. This approach fosters the creation of discriminative representations beneficial for classification and clustering tasks in machine learning. The final loss is thus expressed as:

$$\mathcal{L}_{\text{SCL}} = \sum_{i=1}^{N} \left( -\frac{1}{N_{y_i} - 1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathbb{1}_{y_i=y_j} \log \frac{e^{z(i,j)}}{\sum_{\substack{k=1 \\ k \neq i}}^{N} e^{z(i,k)}} \right) \tag{3}$$

where $N$ is the total number of examples in the minibatch and $N_{y_i}$ represents the number of labels in a mini-batch.

The total loss of the model is therefore the weighted sum of both the cross-entropy loss and the supervised contrastive loss:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{SCL}} + (1 - \lambda) \mathcal{L}_{\text{CEL}} \tag{4}$$

In our experiments, we chose to set $\lambda$ to 0.1 and thus privilege cross entropy loss with a 0.9 weight so that the contrastive loss only acts to consolidate the representation computed through cross entropy.

# 3. Experimental setup

## 3.1. Dataset

We use the ASVspoof 2019 and 2021 datasets for all experiments in this paper [3, 4]. Training is done using the ASVspoof 2019 LA training partition, as it is widely done in the literature and ASVSpoof 2021 does not provide a training set. All models are evaluated on the evaluation parts of ASVspoof 2019 LA, ASVspoof 2021 LA and DeepFake (DF) tasks. It is important to note a massive class imbalance between spoof and bona fide audio (which falls to 2.78% of all audio in DF 2021) as shown in 1. The same can be said of the training data with 10.17% for bona fide audio versus 89.83% for spoof. The spoofed speech employed in both challenges originated from the VCTK database, generated using TTS and VC algorithms in the 2019 edition. To increase the challenge complexity in 2021 edition, the organizers introduced variable transmission and encoding conditions to the synthetic speech. Both the 2019 LA and 2021 LA/DF tasks utilize the same set of speakers.

## 3.2. Model Training

All models were trained using 4 A100 GPUs and a batch size of 50 when using WavLM in evaluation mode. The model used in this study is the pretrained WavLM Large from Huggingface. For the finetuning step, we use a custom data collator that dynamically pads the inputs with the length of the longest input in the batch (batch size of 16).

We used a learning rate of 0.0001, weight decay of 0.0001 and the Adam optimizer. The $\tau$ parameter in the contrastive loss was set to 0.07. The overall architecture is trained with the train subset of ASVspoof 2019 LA database. Given the data imbalance in the dataset, we use weighted cross-entropy loss with higher weight to minority class (bona fide) and lower weight to majority class (spoof). We assigned weights of 0.9 and 0.1 to genuine and spoofed speech, respectively. The models trained for roughly 50 epochs with WavLM evaluation mode. The maximum number of training epochs is 100. Fine-tuning of WavLM layers was performed as joint training with the classifier for several epochs (1-7).

## 3.3. Experiments

In this work, we explore layer-wise feature selection and supervised contrastive learning for WavLM-based spoofed speech detection. We focus our evaluation on two main aspects: the impact of layer selection and the importance of (1) fine tuning and (2) constrastive learning. In this paper, the main metric used for all experiments is the Equal Error Rate (EER). EER corresponds to the point where the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) become equal. A lower EER signifies a more robust and reliable biometric security system.

### *3.3.1. Layer Selection Experiments*

We evaluate two broad categories of layers: the first 12 transformer layers and the last 12 transformer layers. These embeddings are then used to train the classifier. Subsequently, we investigate eight-layer intervals: layers 0-7, 8-15, and 16-24. Each configuration is tested with WavLM in evaluation mode for feature extraction with MLP classifier. This allows us to characterise the potential of each combination for EER. We also add configurations combining the first dozen layers with a selection of top layers as initial results, in line with the literature, indicated that some of the last layers seem to help for speaker identification [15, 25].

### *3.3.2. Fine-tuning and Supervised Contrastive Learning*

For the second part of the study, we first evaluate the impact of fine-tuning. We finetune the classifier jointly with WavLM, freezing the CNN feature extractors and transformer layers deemed less relevant for acoustic information. Cross-entropy loss is employed alone (that is, without the contrastive loss) during fine-tuning for the first experiment.

We then explore the impact of supervised contrastive loss with various hyperparameter configurations for the second experiments including the weighting factor for the contrastive loss $\lambda \in [0.1, 0.3, 0.5, 0.7, 0.9]$ and the temperature parameter $\tau \in [0.07, 0.1, 0.3, 0.5, 0.7]$. $\tau = 0.07$ and $\lambda = 0.1$ were then chosen as it gives the best results. We use these parameters to train the most promising layer configurations: layers 0-12 + 22-23 and layers 2-12 + 22-23.

| WavLM | Layers | Loss | LA 2019 | LA 2021 | DF 2021 |
|---|---|---|---|---|---|
| Eval | 25 | CEL | 0.45 | 4.65 | 5.97 |
| Finetune | | CEL | 0.18 | 3.22 | 6.07 |
| | | CEL + SCL | 0.11 | 4.54 | 4.59 |
| Eval | 0 - 7 | CEL | 0.38 | 11.57 | 10.73 |
| Eval | 8 - 15 | CEL | 0.64 | 4.95 | 6.55 |
| Eval | 16 - 24 | CEL | 1.43 | 6.22 | 9.18 |
| Eval | 0 - 12 | CEL | 0.34 | 7.04 | 7.19 |
| Eval | 13 - 24 | CEL | 1.29 | 5.93 | 9.34 |
| Eval | 0 - 12 + 22 - 23 | CEL | 0.31 | **3.11** | 6.25 |
| Finetune | | CEL | 0.29 | 6.83 | 5.99 |
| | | CEL + SCL | 0.23 | 3.31 | **4.47** |
| Eval | 0 - 12 + 21 - 23 | CEL | 0.53 | 4.23 | 6.38 |
| Eval | 2 - 12 + 21 - 23 | CEL | 3.44 | 3.51 | 6.15 |
| Finetune | | CEL | **0.09** | 4.06 | 4.84 |
| | | CEL + SCL | 0.09 | 6.14 | 5.48 |

Table 2: *Equal Error Rate (EER) values for the different combinations of our model. Eval designates a configuration where only the classifier is trained without joint fine-tuning with WavLM.*

| System | LA 2019 | LA 2021 | DF 2021 |
|---|---|---|---|
| Tak et al. [7] | 1.06 | | |
| Jung et al. [9] | 0.83 | | |
| Chen et al. [32] | 0.58 | | |
| Eom et al. [17] | 0.40 | | |
| Lee et al. [33] | 0.31 | | |
| Wang et al. [34] | 2.31 | 7.18 | 5.44 |
| Wang et al. [34] | 1.28 | 6.53 | 4.75 |
| Martin et al. [20] | | 3.54 | 4.98 |
| Tak et al. [18] | | **0.82** | **2.85** |
| Guo et al. [21] | 0.42 | 2.56 | 5.08 |
| Ours | **0.23** | 3.31 | 4.47 |

Table 3: *EER comparison between our models and other works obtained on LA 2019, LA and DF 2021.*

## 4. Discussion

First, results shown in table 2 seem to confirm that speaker information resides primarily in the lower layers with contributions from a few top layers. For instance, layers 0-12 perform better than 13-24 on LA 2019 and DF 2021. Likewise, layers 8-15 are consistently better than 16-24. But that trend is not general as layers 0-12 perform worse than 13-24 on LA 2021. As suggested in the literature, layers 21 to 23 do provide much improvement when combined to the first layers.

After experimenting on various combinations, best overall results were obtained with layers 0-12 + 22-23 and 2-12 + 21-23. For the latter, even better results were obtained increasing $\lambda$ to 0.9 instead of 0.1: 0.11 (LA 2019), 3.69 (LA 2021) and 5.27 (DF 2021) but this was not the case for other configurations.

Fine-tuning provides mixed results. By extracting speaker-specific features, we finetune the pre-trained WavLM model for the spoofed speech detection task, effectively transferring knowledge learned from massive amounts of data. This approach yielded a significant performance improvement on LA 2019, with an EER of 0.09 for configuration 2-12 + 21-23. However, the model's performance on the ASVspoof 2021 dataset suggests overfitting to the 2019 data and overall, results show a degradation for LA 2021 and DF 2021. The ASVspoof 2019 dataset features high-quality audio without real-world complexities like encoding, transmission effects, or media compression. Consequently, the model struggles to generalize to the more practical scenarios presented in the 2021 LA evaluation set (including telephony) and the diverse compression algorithms and unknown spoofing attacks encountered in the 2021 Deepfake (DF) evaluation set.

When using the contrastive loss, our experiments tend to show more homogeneous results than what we obtain when only using cross entropy. These findings highlight the trade-off between model specificity and generalizability. While layer-wise selection within WavLM offers promise for spoofed speech detection, further research is needed to enhance generalizability to real-world scenarios with diverse audio characteristics and spoofing techniques.

Overall though, our results are competitive with the state of the art as shown on table 3. In particular, our results rank first on LA 2019 while remaining competitive for other datasets, especially DF 2021.

## 5. Conclusion

This work investigates the efficiency of WavLM as a feature extractor and the use of supervised contrastive loss for spoofed speech detection without data augmentation or audio segmentation. We leverage WavLM to extract speaker embeddings for each transformer layer, select layers to be utilised, and use attentive statistical pooling to merge them. While standard fine-tuning enables basic spoofed versus bona fide speech discrimination, employing supervised contrastive learning refines the model's ability to differentiate between the two classes. Our results show that the most informative WavLM layers are found in both the lower and some upper levels of the network. Moreover, while fine-tuning does not provide stable results on all 3 evaluation sets and tends to overfit quickly, the contrastive loss helps achieve much more stable results.

We also suspect that better performance can be reached by enhancing the architecture of our classifier. In particular, paying more attention to the challenge posed by the diversity of bona fide speech in the dataset (or lack thereof) seems promising. Using a more effective training process, implementing dynamic batching for instance, and leveraging data augmentation techniques are also potential leads.

## 6. Acknowledgements

This work was granted access to the HPC/AI resources of IDRIS under the allocation 2023-AD011013889R1 made by GENCI and funded by the Côtes d'Armor departmental council.

## 7. References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech*, 2015.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Interspeech*, 2017.

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

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

[5] T. Sabhanayagam, V. P. Venkatesan, and K. Senthamaraikannan, "A comprehensive survey on various biometric systems," *International Journal of Applied Engineering Research*, 2018.

[6] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, 2018.

[7] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *ASVspoof Challenge*, 2021.

[8] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP*, 2021.

[9] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP*, 2022.

[10] C. Goel, S. Koppisetti, B. Colman, A. Shahriyari, and G. Bharaj, "Towards attention-based contrastive learning for audio spoof detection," in *Interspeech*, 2023.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[12] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP*, 2022.

[13] 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*, 2021.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[15] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, 2022.

[16] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE JSTSP*, 2022.

[17] Y. Eom, Y. Lee, J. S. Um, and H. Kim, "Anti-spoofing using transfer learning with variational information bottleneck," in *Interspeech*, 2022.

[18] 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 *Odyssey*, 2022.

[19] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech*, 2018.

[20] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *ICASSP*, 2022.

[21] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, "Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier," in *ICASSP*, 2024.

[22] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *NeurIPS*, vol. 36, 2023, pp. 19 594–19 621.

[23] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*, 2021.

[24] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP*, 2023.

[25] T. Ashihara, M. Delcroix, T. Moriya, K. Matsuura, T. Asami, and Y. Ijima, "What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis," *arXiv preprint arXiv:2401.17632*, 2024.

[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NeurIPS*, 2020.

[27] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *ICLR*, 2021.

[28] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," in *Interspeech*, 2021, pp. 1194–1198.

[29] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohammed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP*, 2020.

[30] 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 *Interspeech*, 2021.

[31] 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. of ACL-Int. Joint Conf. NLP*, 2021.

[32] F. Chen, S. Deng, T. Zheng, Y. He, and J. Han, "Graph-based spectro-temporal dependency modeling for anti-spoofing," in *ICASSP*, 2023, pp. 1–5.

[33] J. W. Lee, E. Kim, J. Koo, and K. Lee, "Representation selective self-distillation and wav2vec 2.0 feature exploration for spoof-aware speaker verification," in *Interspeech*, 2022, pp. 2898–2902.

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