

How Do Neural Spoofing Countermeasures Detect Partially Spoofed Audio?

Tianchi Liu^{1,2}, Lin Zhang³, Rohan Kumar Das⁴, Yi Ma², Ruijie Tao², Haizhou Li^{5,2}

¹Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR), Singapore

²Department of Electrical and Computer Engineering, National University of Singapore, Singapore

³National Institute of Informatics, Tokyo, Japan ⁴Fortemedia Singapore, Singapore

⁵School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China

liu_tianchi@i2r.a-star.edu.sg, zhanglin@nii.ac.jp, rohankd@fortemedia.com, {mayi, ruijie.tao}@u.nus.edu, haizhouli@cuhk.edu.cn

Abstract

Partially manipulating a sentence can greatly change its meaning. Recent work shows that countermeasures (CMs) trained on partially spoofed audio can effectively detect such spoofing. However, the current understanding of the decision-making process of CMs is limited. We utilize Grad-CAM and introduce a quantitative analysis metric to interpret CMs' decisions. We find that CMs prioritize the artifacts of transition regions created when concatenating bona fide and spoofed audio. This focus differs from that of CMs trained on fully spoofed audio, which concentrate on the pattern differences between bona fide and spoofed parts. Our further investigation explains the varying nature of CMs' focus while making correct or incorrect predictions. These insights provide a basis for the design of CM models and the creation of datasets. Moreover, this work lays a foundation of interpretability in the field of partial spoofed audio detection that has not been well explored previously.

Index Terms: explainable AI, speech anti-spoofing, partial spoof, countermeasures

1. Introduction

The advancement of generative models in speech processing has magnified the threat of spoofed speech. To mitigate the potential damage from the synthetic speech, several advanced CMs have been proposed [1–9]. Most of them focus on the fully spoofed scenario, where the spoofed audio is entirely created by methods such as text-to-speech, voice conversion, and so on.

However, it is not always necessary to generate the entire audio with generative models from the perspective of spoofing. When attackers have access to bona fide samples close to the desired audio, they only need to manipulate arbitrary short parts using generative models. An example is depicted in Figure 1. This has led to a new spoofing scenario referred to as ‘Partial Spoof’ (PS), which has attracted increased attention recently [10–13]. In the PS scenario, when concatenating bona fide and spoofed segments, noticeable artifacts of discontinuity will occur. To address this, several strategies are applied, such as signal-processing techniques [10, 14], neural network-based approaches [15], etc. These strategies aim to mitigate the discontinuity and create smoother transitions. Hence, we describe the resulting segments as transition regions (also known as ‘concatenated parts’ [10]), illustrated in Figure 1.

Given that the spoofed region can be arbitrarily short, detecting spoofed audio in the PS scenario becomes more challenging, especially when training data consists entirely of fully spoofed samples [16]. To improve the performance of detecting partially spoofed audio in the PS scenario, several neural CMs and data sets are proposed [10–13, 17–20]. While these efforts remarkably improve the capability of detecting PS audio, the

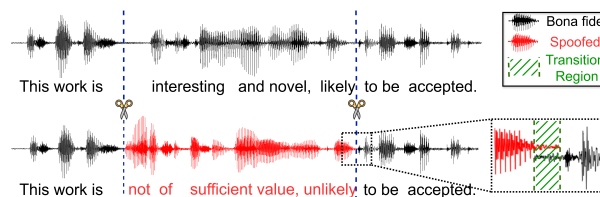


Figure 1: Illustration of partially spoofed speech changing the meaning of a sentence.

inherent characteristics of data-driven networks result in them functioning as *black-boxes* with restricted interpretability.

Understanding CMs is crucial not only for ensuring the trustworthiness of their decisions but also for benefiting the optimization of the model and the design of the training dataset. Therefore, it is essential to understand how CMs work. In light of this, there is a growing interest in the field of *Explainable AI (XAI)* for CMs. The authors of [21] explored the role of the silence region by analyzing its duration for a CM's effectiveness, while the work in [22], analyzed the effect of removing silence and employed gradient-weighted class activation mapping (Grad-CAM) for visualization. Along similar directions, the authors of [23] delved into the analysis of background noise in spoofed speech, whereas the study in [24] focused on analyzing frequency bands and speech formants. The Grad-CAM-Binary example is used in [25] to illustrate that buzziness and rhythm of speech are the determining factors for CMs. Similarly, SHapley Additive exPlanations (SHAP) contributes to exploring the artifacts produced by speech generators [26, 27]. The existing studies mainly center on models trained with entirely spoofed data, but there is very limited exploration regarding partial spoofing CMs. This motivated us to add a new dimension of interpretability to the field of partially spoofed audio detection, an aspect that has not been thoroughly explored yet.

In this work, we aim to understand and explain how partial-spoofing CMs make decisions, thereby providing valuable insights for their development. We consider Grad-CAM [28] which uses the gradients to produce a localization map highlighting the important regions of the input for the model's predictions. While Grad-CAM has been applied in speech-related tasks [22, 25, 29, 30], to the best of our knowledge, this is the first effort to use it for explaining partial-spoofing CMs. We believe that compared to bona fide and spoofed speech parts, the model pays more attention to the artifacts from the transition regions between them. To validate this, we propose a novel metric to quantitatively analyze Grad-CAM scores. Additionally, we investigate the areas that the CM focuses on for the cases of successful and failed detection, offering valuable insights for the refinement of partial-spoofing CMs.

2. Explaining CMs with Grad-CAM

2.1. Grad-CAM



Figure 2: An example of counterfactual explanations with Grad-CAM [28]. Figures are produced using [31].

Grad-CAM is designed to utilize the gradients of a specified target output class, generating a coarse localization map that evaluates each input pixel’s activation values. The larger values indicate that the corresponding regions are more important for the model’s predictions [28].

For an image, assuming $l_k = f(x)$ is the logit of the target class k generated from the pretrained classification model, $R \in \mathbb{R}^{(C \times W \times H)}$ is the hidden representation from the target layer in dimension of *channel* $\times$ *width* $\times$ *height*, the Grad-CAM score S can be defined as follows:

$$S = \text{ReLU}\left(\sum_c R_{c,w,h} \cdot g_{c,w,h}\right), \quad g_{c,w,h} = \frac{\partial l_k}{\partial R_{c,w,h}}. \quad (1)$$

Specifically, the Grad-CAM is the multiplication of the hidden representation and the gradient from the target logits with respect to the hidden representation. The ReLU processing keeps the pixel having positive contributions to the prediction.

An example is illustrated in Figure 2, where Grad-CAM identifies the crucial area in the image with large values of S . This highlights that the area contributes most significantly to the model’s final decision, whether predicting a cat or a dog.

2.2. Explaining CMs with Grad-CAM

We adopt a recent top performing partial-spoofing CM in [10] for analysis. It is structured with a self-supervised learning (SSL) model [32] as front-end and stacked gated multilayer perceptron (gMLP) [33] layers as a classifier, shown in Figure 3 (a). The downsampling block combines max pooling with a kernel size of 2 and a linear layer. The output of SSL-gMLPs in [10] adopts a two-class design of bona fide and spoofing.

To address the *black-box* nature of partial spoofing CMs, we apply Grad-CAM to explain their decision-making process. This method helps analyze the regions that partial spoofing CMs focus on for their detection decisions. As shown in Figure 3 (a), to obtain high-resolution Grad-CAM scores for accurate counterfactual explanations, we select the layer following the SSL front-end as the target layer for Grad-CAM. This choice is due to the model design described in [10], which employs max pooling during the downsampling stage, thereby reducing the time-domain resolution [34].

Considering the linear layers in multiple downsampling blocks and the stacking of multiple gMLPs, the class-discriminative information may become more diluted. This dilution could lead to less accurate or interpretable Grad-CAM scores due to the backward propagation through too many layers. Moreover, the layer immediately following the SSL front-end primarily captures basic speech representations, which may not meaningfully correlate with the CM’s decision-making process in terms of explainability. To address these issues, we optimize the network, as shown in Figure 3 (b), maintaining a high resolution until the last two layers of a pooling and a linear. The SE-Res1D is used to replace original gMLPs, combining a single 1D-ResNet block with squeeze-and-excitation (SE) module [35–37]. The design of this new model aims to overcome the

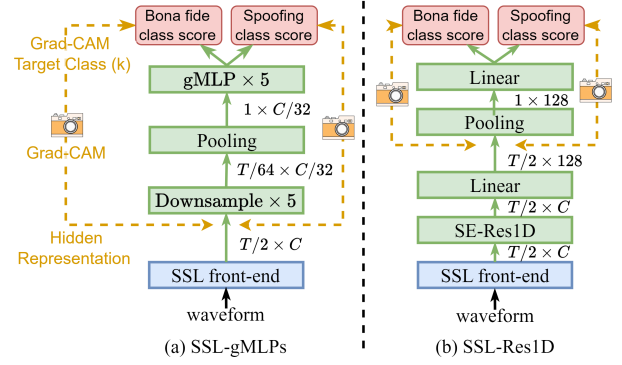


Figure 3: Block diagrams illustrating the structures of the (a) SSL-gMLPs and (b) SSL-Res1D models. The dashed yellow line with a camera icon indicates the Grad-CAM.

above-mentioned two challenges that affect explainability in the original SSL-gMLPs model. This effort leads to an improved version, which we refer to as SSL-Res1D.

Furthermore, in our effort to explain CMs’ decision-making process through Grad-CAM, we aim to obtain accurate and high-resolution Grad-CAM activation values at the frame level for analysis. In this case, we refrain from using any score smoothing techniques or scaling up the frame-level activation scores to match the original waveform’s shape via interpolation.

2.3. How to Analyze Grad-CAM Scores Quantitatively?

Visualizing through Grad-CAM scores, as shown in Figure 2, is highly intuitive but falls short in quantitatively analyzing the entire dataset to draw a dependable conclusion. To quantitatively validate our findings and provide a comprehensive numerical summary of the analysis across the entire dataset, we introduce a metric termed Relative Contribution Quantification (RCQ). The metric is designed to describe the relative relationships between different speech segments concerning their contribution to the final decision.

First, we use the Grad-CAM to derive the frame-level Grad-CAM scores of the i -th sample $\{S_{i,t=1}, S_{i,t=2}, \dots, S_{i,t=T}\}$ in the test set I , corresponding to the input speech sample of frame T . Then we align these scores with the frame-level annotations $y \in \{\text{TR, BS, BN, SS, SN}\}$ which represent five categories: **T**ransition **R**egion, **B**ona fide **S**peech, **B**ona fide **N**on-speech, **S**poofed **S**peech, and **S**poofed **N**on-speech, respectively. For each category, we compute the average Grad-CAM score $\bar{S}$ by averaging Grad-CAM scores across all I samples in the test set. To illustrate, consider the category of transition region (TR):

$$\bar{S}_{\text{TR}} = \frac{1}{F_{\text{TR}}} \sum_{i=1}^I \sum_{t=1}^{T_i} S_{i,t}, \forall i, t \text{ s.t. } y_{i,t} = \text{TR}, \quad (2)$$

where F_{TR} is the total number of frames in the transition region category across the entire test set. In a similar manner, we compute the average Grad-CAM scores for the remaining categories, denoted as $\bar{S}_{\text{BS}}$, $\bar{S}_{\text{BN}}$, $\bar{S}_{\text{SS}}$, and $\bar{S}_{\text{SN}}$. Subsequently, we compute the RCQ for all five categories by evaluating their respective relative values to the average Grad-CAM score of all frames $\bar{S}_{\text{all}}$. The calculation is exemplified for the transition region (TR) category as follows:

$$\text{RCQ}_{\text{TR}} = (\bar{S}_{\text{TR}} - \bar{S}_{\text{all}}) \div \bar{S}_{\text{all}} \times 100\%. \quad (3)$$

The RCQ metric quantitatively evaluates the contribution of each category to the final decision across the entire dataset. It also provides insight into the relative importance of the five categories, where a larger value indicates greater importance.

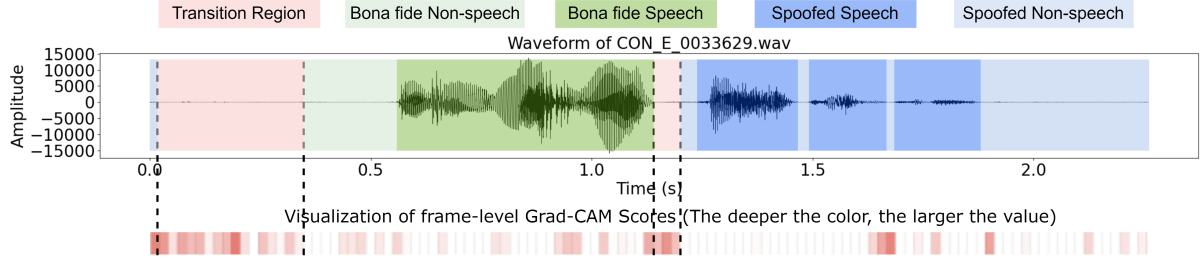


Figure 4: Visualization of the waveform and frame-level Grad-CAM scores for *CON_E_0033629.wav* from the evaluation set.

3. Experimental Setup

3.1. Dataset

In this work, we utilized the public accessible database, PartialSpoof [10], which is notable for providing detailed timestamp annotations. These annotations cover both speech and non-speech parts from bona fide and spoof, as well as transition regions. This comprehensive annotation enables its use in studies aimed at an in-depth understanding of how CMs make decisions.

3.2. Configuration

We mostly follow the study in [10] and its recipe¹ to develop CMs. Specifically, we use the wav2vec2-large [32] as the front-end, *MSE for P2SGrad* loss [3] as the loss function, and Adam as optimizer. To accommodate GPU memory limitations, we reduce the batch size to 2, and adjust the learning rate accordingly to 2.5×10^{-6} . Training is stopped if there is no improvement on the development set for 20 epochs. The optimal epoch on the development set is selected to report the performance and be analyzed by Grad-CAM. The equal error rate (EER) metric is used to report the performance of detecting spoofed audio.

For Grad-CAM, we adopt the PyTorch library for CAM methods [31]², with necessary modification for transitioning from images to speech signals, which involves one dimension less. The frame length for Grad-CAM analysis is 20 ms [10], constrained by the output dimension of wav2vec2-large [32].

4. Results and Discussion

4.1. Partially Spoofed Speech Detection Evaluation

Table 1 reports the performance for the SSL-gMLPs and SSL-Res1D models on the development and evaluation sets of PartialSpoof. A notable performance improvement is observed when utilizing the PartialSpoof dataset for training, compared to using the fully spoofed ASVspoof dataset. Moreover, the SSL-Res1D model shows superior performance in comparison to SSL-gMLPs when trained on the ASVspoof 2019 LA dataset and closely matches the top-tier results in [10] upon training with the PartialSpoof dataset. These results illustrate that the models we analyzed are comparable to those in [10].

4.2. Visualizing Frame-level Grad-CAM

A visualization example of the Grad-CAM score for each frame is shown in Figure 4. As introduced in Section 2.3, partially spoofed audio in the PartialSpoof database comprises five components: the non-speech as well as speech segments of both bona fide and spoofed audio, and overlap-add-based transition region. The Grad-CAM scores for each frame are dis-

Table 1: Performance in EER (%) for SSL-gMLPs and SSL-Res1D, trained on the ASVspoof 2019 LA [38] or PartialSpoof [10] dataset and tested on the development as well as evaluation sets of PartialSpoof. † indicates re-implementation.

Model	Training Set	Dev. Set	Eval. Set
SSL-gMLPs [10]	ASVspoof 2019 LA	12.22	14.19
SSL-gMLPs†	ASVspoof 2019 LA	12.75	13.17
SSL-Res1D (ours)	ASVspoof 2019 LA	4.67	3.60
SSL-gMLPs [10]	PartialSpoof	0.35	0.64
SSL-gMLPs†	PartialSpoof	0.51	0.86
SSL-Res1D (ours)	PartialSpoof	0.35	0.73

played in the lower panel of Figure 4, where a deeper color shade indicates a higher value. These Grad-CAM values signify each frame’s contribution to the final prediction. Observations from Figure 4 reveal that the CM primarily focuses on the transition region considering the evaluation sample *CON_E_0033629.wav*.

4.3. Quantitative Analysis of Grad-CAM Scores

The visualization is intuitive but limited to one example. For quantitative analysis, we apply the RCQ metric from Section 2.3. Our focus is on partially spoofed samples, as bona fide samples in PartialSpoof contain only bona fide parts, excluding spoofed parts and transition regions.

Figure 3 illustrates that each model generates two scores for a sample. These two scores are considered as the probabilities that the utterance is bona fide and spoofed, respectively. Ideally, for a spoofed utterance, the score for the spoof class would be large, whereas the score for the bona fide class would be small. We employ Grad-CAM to separately compute the gradients for these two target classes (k), obtaining the Grad-CAM scores S and deriving RCQs. In this case, we can analyze different focuses of the CMs when predicting scores for these two target classes. The RCQs are reported in Table 2. Below are our discussions and findings:

Differences in focus between CMs trained on fully vs. partially spoofed data: Models trained with ASVspoof and PartialSpoof data show different focuses when detecting partially spoofed audio. The ASVspoof-trained model primarily focuses on spoofed speech when predicting scores for the spoof output class. Conversely, it gives greater emphasis to non-speech, especially bona fide non-speech segments for scoring the bona fide output class. The focus on transition regions, is likely because these regions are predominantly non-speech. In contrast, models trained with PartialSpoof data seem to find an effective solution by focusing mainly on the transition regions.

What is important for CM to learn in PS? The CM trained on the PartialSpoof dataset effectively recognized the transition regions’ pattern, a feature not encountered by the CM trained on the ASVspoof2019 dataset. Linking the difference in

¹<https://github.com/nii-yamagishilab/PartialSpoof>

²<https://github.com/jacobgil/pytorch-grad-cam>

Table 2: *RCQs (%) of SSL-ResID models when predicting spoof and bona fide classes’ scores for partially spoofed samples, across five different segment types. Models are trained on ASVspoof 2019 LA [38] or PartialSpoof [10]. The grey color represents the relative size relationship among the values of the five types of segments within each trial, with deeper shades indicating larger values.*

Training Set	Grad-CAM Target Class (k)	PartialSpoof Development Set						PartialSpoof Evaluation Set					
		Utt.	Bona fide	Spoofed	Transition	Bona fide	Spoofed	Utt.	Bona fide	Spoofed	Transition	Bona fide	Spoofed
		EER	Speech	Speech	Region	Non-speech	Non-speech	EER	Speech	Speech	Region	Non-speech	Non-speech
ASVspoof19	Spoof	4.67	-3.60	31.38	4.07	-29.77	-13.42	3.60	-3.62	25.81	21.69	-33.43	-7.87
	Bona fide		-0.77	-35.07	39.14	30.24	12.75		-7.11	-27.24	37.03	38.72	15.26
PartialSpoof	Spoof	0.35	-0.06	1.87	15.03	-6.95	-1.81	0.73	-0.62	-0.90	25.36	-4.72	0.99
	Bona fide		-3.65	-7.90	49.85	1.27	0.89		-8.63	-4.57	58.01	10.22	-1.09

learned focus to the EER, and considering the identical structure and training strategy, we believe that focusing on transition regions is remarkably effective and efficient when dealing with the PartialSpoof dataset.

Varied focus of CMs in predicting scores for spoof and bona fide classes: Both the ASVspoof2019 and PartialSpoof trained models show noticeably different focuses when predicting spoof and bona fide class scores. The ASVspoof2019 trained model pays more attention to spoofed speech for its spoof class, while prioritizing non-speech for its bona fide class, especially bona fide non-speech. This suggests that the model not only learns the differences in pattern between bona fide and spoofed segments, but also autonomously discerns that speech and non-speech information are respectively suited for different class predictions: non-speech features are more effective for bona fide class predictions, while characteristics of spoofed speech are important for spoof class predictions. Moreover, the ASVspoof2019 trained model strongly emphasizes non-speech over speech when predicting bona fide class, providing insights and evidence into the importance of the non-speech aspect following previous studies [21, 22]. This also highlights the value of this explanatory work. Additionally, the CM trained on the PartialSpoof dataset mostly focuses on the transition regions, this applies to predicting scores for both spoof and bona fide classes. It also exhibits a subtly increased attention to the segments that are particularly relevant to each class.

4.4. What Do CMs Focus on Successful vs. Failed Cases?

Table 3: *RCQs (%) of five different types of speech segments for misclassified partially spoofed samples with SSL-ResID model.*

Test Set	Bona fide Speech	Spoofed Speech	Transition Region	Bona fide Non-speech	Spoofed Non-speech
Dev.	12.22	17.75	-6.76	-19.89	-20.18
Eval.	26.76	27.66	-10.42	-49.19	-28.16

We analyze partially spoofed samples where the PartialSpoof-trained SSL-ResID model made incorrect and correct predictions, employing Grad-CAM with the spoof class as the target class.

Table 3 shows RCQs when the model makes incorrect predictions. It shows that the model’s RCQ values for transition regions are much lower than for bona fide or spoofed speech. Comparing Table 3 with Table 2 of the same PartialSpoof-trained model, we notice that the model mainly pays attention to speech segments and neglects the transition regions when it classifies wrongly. This suggests that the model’s inability to accurately focus on transition regions remarkably contributes to its mistaken identification of spoofed samples.

In addition to analyzing failure cases, we investigate how the CM’s utterance-level prediction scores correlate with the

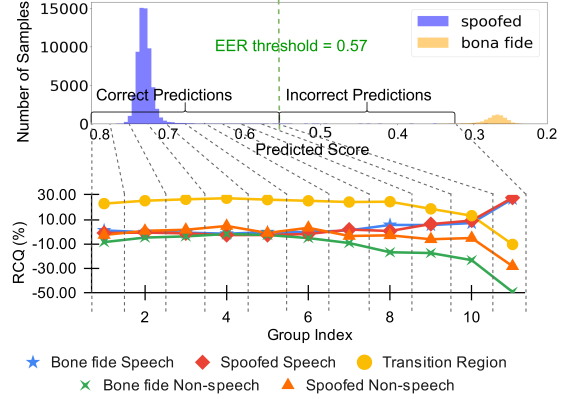


Figure 5: *The prediction score distribution of SSL-ResID trained on PartialSpoof and tested on the evaluation set, along with the line chart showing the RCQ of the five types of segments across 11 sample groups of partially spoofed samples.*

focus areas in correctly predicted samples, as depicted in Figure 5. In the upper panel of Figure 5, we equally divide the range between the maximum and minimum predicted scores of the correctly predicted samples into ten groups. The 11-th group consists of incorrectly classified samples. For each group, we analyze the focus areas of the CM using the RCQ, shown in the lower panel of Figure 5. We notice that correctly classified cases in the model’s predictions are linked to major attention on the transition regions. Conversely, as the output scores near the threshold, the model shifts its focus from the transition region to the speech parts. This trend highlights the importance of transition regions in the CM’s performance, aligning with the aforementioned observation in Section 4.3.

5. Conclusion

In this study, we use Grad-CAM to explain how neural CMs detect partially spoofed audio. We observe that the partially spoofed data-trained CMs not only focus on the spoofed and bona fide areas, but also significantly on the transition regions created by the overlap-add operation during the dataset creation. We introduce a quantitative analysis metric, namely RCQ, to validate this observation across the entire evaluation set. Moreover, we conduct an analysis comparing correctly and incorrectly classified samples. We find that the CM pays less attention to the transition regions when it classifies incorrectly. In contrast, for the samples that are classified correctly, there is a higher focus on the transition regions for samples away from the EER threshold, confirming the important role of these transition regions in the effectiveness of the partial spoof CMs. These insights offer valuable contributions to the development of partial spoofing CMs and the creation of related datasets.

6. Acknowledgements

This work is supported by the National Research Foundation, Prime Minister's Office, Singapore, and the Ministry of Communications and Information, under its Online Trust and Safety (OTS) Research Programme (MCI-OTS-001); Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002); Shenzhen Science and Technology Program ZDSYS20230626091302006.

7. References

- [1] Z. Wu, N. Evans, T. Kinnunen, J. Y. and Federico Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.
- [2] 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.
- [3] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Interspeech*, 2021, pp. 4259–4263.
- [4] H. Tak, M. Todisco, X. Wang, J. weon 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.
- [5] X. Liu *et al.*, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.
- [6] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.
- [7] 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 *Proc. ICASSP*, 2022, pp. 6367–6371.
- [8] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," in *Proc. Interspeech*, 2023, pp. 4009–4013.
- [9] Y. Zang, Y. Zhang, and Z. Duan, "Phase perturbation improves channel robustness for speech spoofing countermeasures," in *Proc. Interspeech*, 2023, pp. 3162–3166.
- [10] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023.
- [11] J. Yi *et al.*, "ADD 2023: The second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.
- [12] Z. Cai, W. Wang, and M. Li, "Waveform boundary detection for partially spoofed audio," in *Proc. ICASSP*, 2023, pp. 1–5.
- [13] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "MLAAD: The multi-language audio anti-spoofing dataset," *arXiv preprint arXiv:2401.09512*, 2024.
- [14] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-Truth: A partially fake audio detection dataset," in *Proc. Interspeech*, 2021, pp. 1654–1658.
- [15] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, and K. Stefanov, "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset," *arXiv preprint arXiv:2311.15308*, 2023.
- [16] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, "An initial investigation for detecting partially spoofed audio," in *Proc. Interspeech*, 2021, pp. 4264–4268.
- [17] J. M. Martín-Doñas and A. Álvarez, "The Vicomtech audio deepfake detection system based on wav2vec2 for the 2022 ADD challenge," in *Proc. ICASSP*, 2022, pp. 9241–9245.
- [18] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.-M. Wang, and H. Meng, "Partially fake audio detection by self-attention-based fake span discovery," in *Proc. ICASSP*, 2022, pp. 9236–9240.
- [19] Z. Cai and M. Li, "Integrating frame-level boundary detection and deepfake detection for locating manipulated regions in partially spoofed audio forgery attacks," *Computer Speech & Language*, vol. 85, p. 101597, 2024.
- [20] J. Yi *et al.*, "ADD 2022: the first audio deep synthesis detection challenge," in *Proc. ICASSP*, 2022, pp. 9216–9220.
- [21] N. Müller, F. Dieckmann, P. Czepin, R. Canals, K. Böttinger, and J. Williams, "Speech is Silver, Silence is Golden: What do ASVspoof-trained models really learn?" in *Proc. ASVspoof Challenge workshop*, 2021, pp. 55–60.
- [22] Y. Zhang, Z. Li, J. Lu, H. Hua, W. Wang, and P. Zhang, "The impact of silence on speech anti-spoofing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3374–3389, 2023.
- [23] D. Salvi, T. S. Balcha, P. Bestagini, and S. Tubaro, "Listening between the lines: Synthetic speech detection disregarding verbal content," *arXiv preprint arXiv:2402.05567*, 2024.
- [24] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification," in *Proc. Odyssey*, 2020, pp. 333–340.
- [25] B. Halpern, F. Kelly, R. van Son, and A. Alexander, "Residual networks for resisting noise: Analysis of an embeddings-based spoofing countermeasure," in *Proc. Odyssey*, 2020, pp. 326–332.
- [26] W. Ge, J. Patino, M. Todisco, and N. Evans, "Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations," in *Proc. ICASSP*, 2022, pp. 6387–6391.
- [27] W. Ge, M. Todisco, and N. Evans, "Explainable deepfake and spoofing detection: An attack analysis using SHapley Additive exPlanations," in *Proc. Odyssey*, 2022, pp. 70–76.
- [28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017, pp. 618–626.
- [29] P. Li, T. Wang, L. Li, A. Hamdulla, and D. Wang, "How phonemes contribute to deep speaker models?" *arXiv preprint arXiv:2402.02730*, 2024.
- [30] Y. Ma, K. A. Lee, V. Hautamäki, M. Ge, and H. Li, "Gradient weighting for speaker verification in extremely low signal-to-noise ratio," *Proc. ICASSP*, 2024.
- [31] J. Gildenblat and contributors, "Pytorch library for cam methods," <https://github.com/jacobgil/pytorch-grad-cam>, 2021.
- [32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 12 449–12 460.
- [33] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 9204–9215.
- [34] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Golden gemini is all you need: Finding the sweet spots for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2324–2337, 2024.
- [35] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
- [36] T. Liu, R. K. Das, K. A. Lee, and H. Li, "MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," in *Proc. ICASSP*, 2022, pp. 7517–7521.
- [37] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Disentangling voice and content with self-supervision for speaker recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 50 221–50 236.
- [38] X. Wang *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech and Language*, vol. 64, p. 101114, 2020.