Interspeech 2024
1-5 September 2024, Kos, Greece

Spoof Diarization: “What Spoofed When” in Partially Spoofed Audio

Lin Zhang[1,2], *Xin Wang*[1], *Erica Cooper*[1,3], *Mireia Diez*[4], *Federico Landini*[4], *Nicholas Evans*[5], *Junichi Yamagishi*[1,2]

[1]National Institute of Informatics, Tokyo, Japan [2]SOKENDAI, Kanagawa, Japan
[3]National Institute of Information and Communications Technology, Kyoto, Japan
[4]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[5]Digital Security Department, EURECOM, France

{partialspoof, lzhang.as}@gmail.com {wangxin, jyamagis}@nii.ac.jp

Abstract

This paper defines **Spoof Diarization** as a novel task in the Partial Spoof (PS) scenario. It aims to determine *what spoofed when,* which includes not only locating spoof regions but also clustering them according to different spoofing methods. As a pioneering study in spoof diarization, we focus on defining the task, establishing evaluation metrics, and proposing a benchmark model, namely the Countermeasure-Condition Clustering (3C) model. Utilizing this model, we first explore how to effectively train countermeasures to support spoof diarization using three labeling schemes. We then utilize spoof localization predictions to enhance the diarization performance. This first study reveals the high complexity of the task, even in restricted scenarios where only a single speaker per audio file and an oracle number of spoofing methods are considered. Our code is available at `https://github.com/nii-yamagishilab/PartialSpoof`.

Index Terms: partial spoof, spoof diarization, countermeasure, clustering

1. Introduction

The Partial Spoof (PS) scenario has recently drawn increasing attention [1, 2]. In the conventional fully spoofed scenario, the entire audio signals – typically an *utterance* within the speech spoofing community – are generated through Text-to-speech (TTS) and/or Voice Conversion (VC) algorithms [3, 4]. In the PS scenario, partially spoofed audio contains segments generated by TTS/VC, with the remaining regions originating from real human speech [1], where all these segments can have varying durations. Compared with the conventional fully spoofed scenario, the PS scenario is more realistic and threatening. Attackers may not need to construct an entire spoofed audio to achieve their goals. Instead, it is more efficient to manipulate only a few arbitrary, short parts of an audio file to drastically distort its original meaning with phonology knowledge and advanced generative models [5].

A number of studies have addressed the PS scenario. Some investigate spoof detection [2, 5, 6, 7] to detect whether an audio is spoofed, while others explored spoof localization [8, 9, 10] to identify specific suspect segments within an otherwise bona fide audio file. However, these two tasks only focus on the binary (yes/no) question to answer whether an audio or segment is spoofed. Thus, solutions for these tasks may be insufficient for realistic forensic scenarios when it is crucial to know not only whether the audio is spoofed but also to obtain detailed information about spoofed segments, like specific spoofing methods or other cues from the spoofing algorithms. This information could potentially aid in tracing the origin or creator of the spoof. We call this task “Spoof Diarization.” When the partially spoofed audio contains segments that are created using multiple generative models or through a recursive generation process, spoof diarization can help with traceability, which is crucial in a court of law. Similar investigations have been conducted in the tampering field to trace devices used for recording [11, 12]. In contrast, binary-classification-based spoof detection and localization lack direct traceability for such purposes. Therefore, it is essential to explore the question of *what spoofed when.*

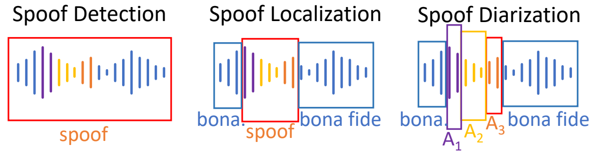


Figure 1: *Spoof detection, localization, and diarization.*

Speaker diarization [13, 14] is a popular task in the speech processing field, aiming to determine “who spoke when” for a given recording. Each recording usually involves an unknown number of speakers whose speech duration varies, as seen in interviews, meetings, broadcasts, etc. In speaker diarization, the input audio file is divided into speaker-homogeneous regions (turns), which are clustered into different groups according to speaker characteristics in those regions. Note that speakers' turns might overlap, with systems expected to correctly label such situations as well.

Aligned with speaker diarization, we define the task “Spoof Diarization” which involves not only locating spoof regions but also clustering them according to spoofing methods. Unlike speaker diarization where clustering is performed according to speaker characteristics, clustering in spoof diarization depends on the spoofing methods. Meanwhile, spoof diarization considers two primary groups of clusters: bona fide and spoof.

In this paper, we focus on defining the task and its evaluation metrics. Further, we present a benchmark model and analyze different labeling schemes corresponding to the specificities of the task.

2. Spoof Diarization

2.1. Definition

To “diarize” means to make a note or keep track of an event in a diary [13]. “Spoof” indicates falsely claiming a speaker identity [15]. Here we define “Spoof Diarization” to annotate the spoofing events. Spoof diarization aims to address *“what spoofed when”* in an audio that contains an unknown number of spoofed segments with variable durations. It is related to but different from two well-known tasks in the PS scenario, spoof detection [2, 5] and spoof localization [1, 8, 10], as shown in Fig. 1. Spoof detection aims to determine whether an audio

10.21437/Interspeech.2024-1365

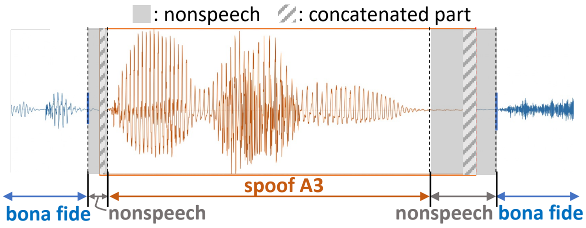
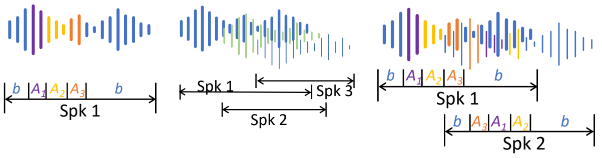


Figure 2: *Example of annotated class-homogeneous segments within an audio in the PS scenario.*



(a) *Spoof Diarization* (b) *Speaker Diarization* (c) *Spoof-Speaker Diarization*

Figure 3: *Comparison of different diarization tasks. "b" for bona fide, "A_*" for spoofing methods, and "Spk*" for speakers. Nonspeech is omitted for clarity.*

contains any spoofed segment. Spoof localization aims to locate spoof and bona fide regions within an audio. Spoof diarization can be viewed as an extension of spoof localization, where segments generated by different spoofing methods are distinguished and assigned different labels. It can be formulated as:

- **Spoof Diarization:** learn a function f_{dia} that takes an audio input $\boldsymbol{x}_{1:T}$, and produces a sequence of multi-class labels $\boldsymbol{c}_{1:M}$:

$$f_{\text{dia}}: \quad \boldsymbol{x}_{1:T} \mapsto \boldsymbol{c}_{1:M},$$

$$c_m \in \{bona \quad fide, A_1, \cdots, A_N, [ConP]\}.$$

Here, $\boldsymbol{x}_{1:T}$ denotes a waveform with T samples, and $\boldsymbol{c}_{1:M}$ denotes the segment-level predictions for M segments. Segments may vary or be uniform in duration, depending on the model, with the minimum duration being a frame (as in this paper). Fig. 2 shows an example annotation. A_* denotes different types of spoofing methods. Note that, during prediction, N is expected to be unknown in real-world scenarios, with a majority of A_* unseen in training data, posing an "open-set" challenge. $ConP$ represents concatenated parts where segments with different classes are seamlessly joined. It can be implemented by signal-processing techniques [1, 8], neural network-based approaches [16], etc., which could also introduce artifacts and could be treated as a special type of spoof. Whether to include it depends on the data design and model implementation.

The ideal spoof diarization system should be able to classify spoofing methods seen in the train set and group unseen methods as in object detection and discovery [17]. As an initial study on this topic, we cluster (without identifying) all segments as done in speaker diarization, and all A_*, seen or unseen during training, will be considered independently on evaluation. Addressing the identification of the exact spoofing methods, especially in an open-set scenario, is designated for future work.

2.2. Spoof diarization and speaker diarization

In the speech processing field, one of the well-known diarization problems is speaker diarization, which aims to determine "who spoke when." A comparison of Fig. 3a (spoof diarization) and 3b (speaker diarization) shows similarities and differences between the two tasks:

- **Similarities:**
  1. Audio samples in both spoof diarization and speaker diarization are generated by an unknown number of classes (spoofing methods or speakers),
  2. Class-homogeneous regions in both tasks can have variable durations.
- **Differences:**
  1. *Duration of turns*: The relevance of detecting short (word level) turns in speaker diarization systems depends on the application and is not relevant or even not evaluated for some of them. In contrast, the detection of such common short-turn spoofed speech (a single word or even a single phoneme) is crucial, as it can completely change the meaning of the audio. e.g., "lock account" to "unlock account."
  2. *Two primary groups of clusters*: In speaker diarization, speakers may vary for each audio. However, in spoof diarization, two primary groups should be considered: bona fide and spoof.

Spoof-speaker diarization, as shown in Fig. 3c, extends spoof diarization to the case of an audio with multiple speakers. This study focuses on spoof diarization, and multi-speaker partially spoofed audio is left for future work.

2.3. Metric - Spoof Jaccard error rate

In speaker diarization, there are two main metrics: diarization error rate (DER) [18], and Jaccard error rate (JER) [19]. While DER accounts for all errors with respect to the total duration of speech in recordings, JER gives equal weight to all speakers, independently of their relative activity. Given the nature of spoof diarization, where a very short segment of spoofed speech can have a large impact, the JER metric matches the task better.

Spoof diarization has two important goals: (1) differentiating spoofed from bona fide segments, and (2) discriminating different spoofing methods. To measure the performances regarding these two goals, we adapted the $\texttt{JER}$[1] with two forms: $\texttt{JI}_{\text{bona}}$[2] and $\texttt{JER}_{\text{spoof}}$, respectively.

They are calculated after an optimal mapping between the reference class and the predicted cluster. This mapping can be determined by the Hungarian algorithm [20] following the common approach in speaker diarization [21, 22]. An example after best mapping is shown in Fig. 4c. $\texttt{JI}_{\text{bona}}$ and $\texttt{JER}_{\text{spoof}}$ for the j-th audio are given by:

$$\texttt{JI}_{\text{bona},j} = \frac{\texttt{FA}_{\text{bona},j} + \texttt{MD}_{\text{bona},j}}{\texttt{TOTAL}_{\text{bona},j}}, \tag{1}$$

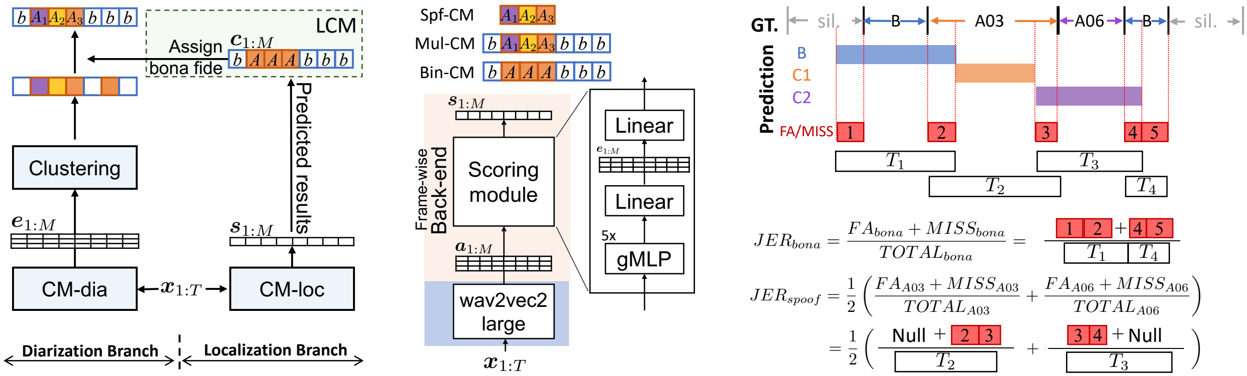
$$\texttt{JER}_{\text{spoof},j} = \frac{1}{|\mathcal{A}_j|} \sum_{A_i \in \mathcal{A}_j} \texttt{JER}_{A_i,j} = \frac{1}{|\mathcal{A}_j|} \sum_{A_i \in \mathcal{A}_j} \frac{\texttt{FA}_{A_i,j} + \texttt{MD}_{A_i,j}}{\texttt{TOTAL}_{A_i,j}}. \tag{2}$$

Where $\texttt{FA}_j$ and $\texttt{MD}_j$ present false alarms and missed detections in the j-th trial, respectively. $\texttt{TOTAL}_j$ refers to the union duration between reference and prediction. Subscripts $_{\text{bona}}$ and $_{A_i}$ indicate bona fide and a specific spoofing method A_i, respectively. $\mathcal{A}_j$ is the set of different spoofing methods within the j-th audio, and $|\cdot|$ denotes the size of the set.

For the global evaluation of a set of audio files $\mathcal{D}$, we use macro averaging:

[1] `https://github.com/nryant/dscore`

[2] JI is used instead of JER because it only considers a single class - bona fide, which aligns better with the Jaccard Index.



(a) *3C model with two branches.* (b) *Structure of CMs in (a).* (c) $\mathtt{JI}_{bona}$ *and* $\mathtt{JER}_{spoof}$ *for spoof diarization.*

Figure 4: *Proposed benchmark model and metrics for spoof diarization.*

$$\mathtt{JI}_{\text{global\_bona}} = \frac{1}{|\mathcal{D}|}\sum_{j\in\mathcal{D}} \mathtt{JI}_{\text{bona},j} \tag{3}$$

$$\mathtt{JER}_{\text{global\_spoof}} = \frac{1}{\sum_{j\in\mathcal{D}}|\mathcal{A}_j|}\sum_{j\in\mathcal{D}}\sum_{A_i\in\mathcal{A}_j}\mathtt{JER}_{A_i,j} \tag{4}$$

To simplify the notation, we drop $_{\text{global}}$ from the subscript in the rest of this paper, e.g., $\mathtt{JI}_{\text{global\_bona}}$ becomes $\mathtt{JI}_{\text{bona}}$.

3. Proposed 3C Model for Spoof Diarization

3.1. 3C model: CM-condition clustering

The proposed 3C model is depicted in Fig. 4a, with the structure of its CMs shown in (b). The 3C model comprises two branches: (1) The diarization branch parallels the conventional modular-based speaker diarization pipeline. It consists of an embedding extractor (CM-dia) followed by a clustering. And (2) the localization branch provides additional bona fide information derived from CM-loc to the diarization branch. CMs in both branches use frame-wise uniform segmentation within audio to facilitate accurate diarization.

To formulate the processing, given an input $\boldsymbol{x}_{1:T}$, we first extract embeddings $\boldsymbol{e}_{1:M}$ from CM-dia. Following this, we calculate a pairwise affinity matrix which will be used for further clustering. From CM-loc, we derive frame-level predicted scores $\boldsymbol{s}_{1:M}$ to determine bona fide frames. It can be implemented by either binary classification or multi-classification strategies. For the binary classification, frame-wise equal error rate [23] is estimated on the development set, and its corresponding threshold τ is used to classify the i-th frame as bona fide if $s_i > \tau$, otherwise as spoofed. In the case of a multi-classification, where training labels include bona fide and various spoofing methods, the class receiving the highest predicted probability is determined for each frame. The bona fide predictions obtained with binary or multi-class CM-loc are then used to potentially enhance the accuracy of bona fide prediction. We use a Label-based CM-constraint (LCM) approach: if CM-loc identifies a frame as bona fide, the final output for this frame is bona fide, (conditioning the CM-dia output). We also explored the potential of using predicted scores to refine the pairwise affinity matrix, but no notable improvement compared with LCM was observed.

3.2. Labeling scheme

As introduced in Section 2.2, in spoof diarization we aim to distinguish bona fide and different spoofing methods. Therefore, the natural choice for training a spoof diarization method, would be to do it for multi-classification considering the labels of all classes. Still, in the PS scenario, it is more relevant to distinguish the two main categories, namely *bona fide* and *spoof*. Taking this into consideration, and given that the 3C model proposed for the spoof diarization contains two branches, we explored the possibilities of training each of these branches with different labeling schemes:

1. `Bin-CM`: CM trained for binary-classification with labels $y_m \in \{bona\ fide, spoof\}$,
2. `Mul-CM`: CM trained for multi-classification with labels $y^*_m \in \{bona\ fide, A_1, \cdots A_N, ConP\}$,
3. `Spf-CM`: CM trained for multi-spoof classification with $y^\star_m \in \{A_1, \cdots, A_N, ConP\}$.

Where `Bin-CM` focuses solely on binary classification between bona fide and spoof, and aggregates all different spoofing methods as a single *spoof* category following previous spoof detection and localization studies. `Mul-CM` not only discriminates between spoof and bona fide but also identifies different spoofing methods. Finally, `Spf-CM`, by excluding bona fide data in its training, aims to explore whether such an approach enhances the ability to differentiate among various spoofing methods.

4. Experiments and Results

4.1. Experimental setup

We used the PartialSpoof[3] database to explore spoof diarization. In this dataset, the proportions of bona fide are 55.3%, 56.0%, and 60.7% for train, development, and evaluation sets, respectively. Furthermore, the presence of different spoofing methods in the PartialSpoof is relatively even and shows a balanced distribution. We applied an adapted Oracle voice activity detection (VAD). Specifically, we removed the nonspeech parts, apart from concatenated parts[4], from training. Nonspeech segments are not taken into account when scoring the predictions. Therefore, errors 1 and 5 shown in Fig. 4c will not happen.

Given the necessity for short-duration diarization highlighted in Section 2.2 – Difference 1, all CMs in this paper were trained at a 20 ms resolution following [1]. They consist of a wav2vec2-large [25] as the front-end with gMLPs [26] as the back-end and achieve the best performance on the spoof localization. For the binary classification CM-loc, 10 ms frame-wise

[3] `https://zenodo.org/records/5766198`

[4] While these concatenated parts are originally nonspeech regions within the PartialSpoof database, modifications made to them might have introduced artifacts [1]. And such concatenated parts are proven helpful for spoof detection [24]. Therefore, we kept and treated them as a unique class during training.

equal error rate [23] was calculated to determine τ for identity whether a frame is bona fide or spoof. The frame-wise embeddings $\boldsymbol{e}_{1:M}$ and predicted scores $\boldsymbol{s}_{1:M}$ are extracted per 20 ms. The resolution of 20 ms is given by convolutional layers in the wav2vec2-large model. The embeddings $\boldsymbol{e}_{1:M}$ are extracted from the penultimate layer of the back-end as shown in Fig. 4b. We utilize the widely used Agglomerative Hierarchical Clustering with cosine distance to cluster the extracted embeddings.

Note that in spoof diarization, the number of spoofing method types can be either known or unknown. In this initial study, we cluster until reaching the oracle number of clusters. In the PS scenario, both spoof and bona fide regions can be arbitrarily short and very difficult to discriminate. This makes accurately estimating the correct number of clusters a complicated task. Therefore, this study conducts an analysis in a controlled scenario with a known oracle number of clusters. Future research could fruitfully explore efficient and accurate methods for estimating cluster numbers in partially spoofed audio.

4.2. How to train CMs to support spoof diarization

4.2.1. How do labeling schemes affect the ability of CMs?

As introduced in Sec. 3.2, there are three possible labeling schemes to train CMs. In this subsection, we focus on understanding their impact on CMs for spoof diarization. First, we exclude the localization branch from the 3C model and train only the diarization branch based on the CMs using three different labeling schemes: `Bin-CM`, `Mul-CM`, and `Spf-CM`. Results are shown in the top part of Table 1.

Looking at the results on the development set, the model trained with `Mul-CM` obtains overall best results. This could be expected, as it is the only labeling scheme that covers all the classes. However, it is somewhat surprising to see that it performs similarly to `Bin-CM` for bona fide localization, even if the latter was specifically trained only for this task. Regarding the model `Spf-CM`, one could expect poor performance on for bona fide location, but the model also underperforms in terms of $\mathtt{JER}_{\text{spoof}}$, revealing that specific treatment of the bona fide class is needed during system training.

The remarkably higher $\mathtt{JI}_{\text{bona}}$ and $\mathtt{JER}_{\text{spoof}}$ observed across all three CMs in the evaluation set is understandable, given that the evaluation set of the PartialSpoof database contains seven out of thirteen spoofing methods that are completely unseen during training [3], thereby presenting a more complex challenge. Focusing on such evaluation set results, we can still see some interesting patterns: the `Mul-CM` suffers higher degradation in JI_{bona} compared to `Bin-CM` on the evaluation set, which might indicate overfitting to the development set.

4.2.2. How do we utilize CMs trained under varying labeling schemes?

Based on the analysis in the previous section, we first chose `Mul-CM` as CM-dia to produce embeddings for various spoofing methods. Besides, considering that the task of locating bona fide segments is now part of the localization branch, we still considered `Spf-CM` to generate the embeddings for clustering. Moreover, considering that both `Mul-CM` and `Bin-CM` showed effective performance in locating bona fide regions, we evaluated these two models as potential options for the CM-loc. The bottom of Tab. 1 shows results for our proposed 3C model.

First, we analyze the efficacy of the proposed 3C model. The Dia-`Mul-CM` models with and without localization branch perform similarly on the development set. However, in the evaluation set, we observe a decrease in $\mathtt{JI}_{\text{bona}}$ but an increase in $\mathtt{JER}_{\text{spoof}}$. That is, integrating the localization branch allows for better localization capabilities but it negatively impacts the efficiency of diarizing spoofing methods. When comparing Dia-`Spf-CM` before and after introducing the localization branch, we notice not only an (expected) remarkable improvement in $\mathtt{JI}_{\text{bona}}$, but also an improvement in $\mathtt{JER}_{\text{spoof}}$ in both the development and evaluation sets. Still, Dia-`Mul-CM` performs better than Dia-`Spf-CM`, which reveals that the Dia-`Mul-CM` model extracts better embeddings for diarizing spoofing methods. Therefore, model selection would depend on the relevance of these two goals for the specific use case.

Table 1: *Results on the PartialSpoof database. Confidence intervals within "()" were calculated using the Interspeech official toolkit with the default configuration. "`-CM`" in the first two columns are omitted.*

Model		Development set		Evaluation set	
Dia.	Loc.	$\mathtt{JI}_{\text{bona}}$ (%)	$\mathtt{JER}_{\text{spoof}}$ (%)	$\mathtt{JI}_{\text{bona}}$ (%)	$\mathtt{JER}_{\text{spoof}}$ (%)
`Bin`	/	4.37 (±0.06)	20.45 (±0.34)	16.85 (±0.14)	33.13 (±0.22)
`Mul`	/	4.49 (±0.07)	5.21 (±0.11)	19.66 (±0.14)	28.05 (±0.17)
`Spf`	/	26.17 (±0.23)	20.85 (±0.25)	32.30 (±0.16)	38.51 (±0.17)
`Mul`	`Bin`	4.49 (±0.07)	5.27 (±0.11)	15.15 (±0.12)	34.13 (±0.19)
	`Mul`	4.59 (±0.07)	5.31 (±0.10)	17.08 (±0.12)	35.34 (±0.19)
`Spf`	`Bin`	4.52 (±0.07)	5.71 (±0.12)	15.18 (±0.11)	36.03 (±0.19)
	`Mul`	4.62 (±0.08)	5.81 (±0.12)	17.10 (±0.12)	37.78 (±0.18)

Second, comparing the different CM-loc options, results show that using `Bin-CM` as CM-loc slightly outperforms using `Mul-CM` as CM-loc, which aligns with the observation from the previous subsection. That is expected as `Bin-CM` is specialized in distinguishing bona fide from spoof.

As already pointed out, performance on $\mathtt{JER}_{\text{spoof}}$[5] is poor on the evaluation set, mainly due to the complexity posed by the open-set scenario on the spoof diarization task. To get a better insight into the impact that unknown spoofing methods had on performance, we analyzed how the best model, namely Dia-`Mul-CM` + Loc-`Bin-CM`, performed in terms of $\mathtt{JER}_{\text{spoof}}$ in known (11.98%) and unknown (49.02%) spoofing methods[6] separately. Results show a considerable performance gap between them which will be addressed in future work.

5. Conclusion

Using generative models to manipulate arbitrary short segments of an audio can drastically change its original meaning as in the Partial Spoof scenario. Locating and tracing back such partially spoofed segments is crucial for speech security, like in a court of law. Thus, spoof diarization, aiming to answer *what spoofed when*, is an essential task that should be explored for the PS scenario.

This study serves as a foundational exploration, presenting the task definition, evaluation metrics, and a benchmark model. Preliminary results in this first study indicate the high complexity of the task, even in controlled scenarios with just a single speaker per audio and a predetermined oracle number of spoofing methods. We hope that our insights can inspire and encourage future investigations into this task.

[5] A breakdown of $\mathtt{JER}_{\text{spoof}}$ for the individual spoofing methods in the evaluation set can be found in the appendix of the arXiv version: `https://arxiv.org/abs/2406.07816`

[6] Known and unknown are grouped by the presence of their acoustic model and waveform generator in the training dataset, following ASVspoof 2019 LA database [27].

6. Acknowledgements

This study is partially supported by JST AIP Acceleration Research (JPMJCR24U3), JST CREST Grant (JPMJCR20D3), JST the establishment of university fellowships Grant (JPMJFS2136), SOKENDAI Student Dispatch Program, and MEXT KAKENHI Grants (21H04906). Mireia Diez and Federico Landini from Brno University of Technology were supported by Czech Ministry of Interior projects Nos. VJ01010108 "ROZKAZ". Nicholas Evans from EURECOM was supported by the ANR in France (BRUEL project, ANR-22-CE39-0009).

7. References

[1] 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.

[2] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "Add 2022: the first audio deep synthesis detection challenge," in *Proc. ICASSP*, 2022, pp. 9216–9220.

[3] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech and Language*, vol. 64, p. 101114, 2020.

[4] 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 Workshop*, 2021, pp. 47–54.

[5] 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.

[6] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *Proc. ICASSP*. IEEE, 2022, pp. 9241–9245.

[7] 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*. IEEE, 2022, pp. 9236–9240.

[8] 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*, 2021, pp. 1654–1658.

[9] L. Zhang, X. Wang, E. Cooper, and J. Yamagishi, "Multi-task Learning in Utterance-level and Segmental-level Spoof Detection," in *Proc. ASVspoof Workshop*, 2021, pp. 9–15.

[10] B. Zhang and T. Sim, "Localizing fake segments in speech," in *Proc. ICPR 2022*. IEEE, 2022, pp. 3224–3230.

[11] L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *Proc. MMSP*, 2013, pp. 177–182.

[12] D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth, and S. Tubaro, "Audio splicing detection and localization based on acquisition device traces," *IEEE Transactions on Information Forensics and Security*, 2023.

[13] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 356–370, 2012.

[14] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[15] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

[16] 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.

[17] J. Zheng, W. Li, J. Hong, L. Petersson, and N. Barnes, "Towards open-set object detection and discovery," in *Proc. CVPR Workshops*, 2022, pp. 3960–3969.

[18] J. Fiscus, J. Ajot, M. Michel, and J. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation." Rich Transcription Spring Meeting Recognition Evaluation 2006, 2006-05-04 2006.

[19] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, task, and baselines," *Proc. Interspeech*, 2019.

[20] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[21] H. Bredin, " pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems," in *Proc. Interspeech*, 2017, pp. 3587–3591.

[22] "Nist rich transcription evaluations," https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation, version: md-eval-v22.pl.

[23] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "Range-based equal error rate for spoof localization," in *Proc. Interspeech*, 2023, pp. 3212–3216.

[24] Z. Cai, W. Wang, and M. Li, "Waveform boundary detection for partially spoofed audio," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[25] 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.

[26] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," in *Proc. NeurIPS*, vol. 34, 2021, pp. 9204–9215.

[27] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.