INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Range-Based Equal Error Rate for Spoof Localization

Lin Zhang[1,2], *Xin Wang*[1], *Erica Cooper*[1], *Nicholas Evans*[3], *Junichi Yamagishi*[1,2]

[1]National Institute of Informatics, Tokyo, Japan
[2]SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
[3]Digital Security Department, EURECOM, France

{zhanglin, wangxin, ecooper, jyamagis}@nii.ac.jp, evans@eurecom.fr

Abstract

Spoof localization, also called segment-level detection, is a crucial task that aims to locate spoofs in partially spoofed audio. The equal error rate (EER) is widely used to measure performance for such biometric scenarios. Although EER is the only threshold-free metric, it is usually calculated in a point-based way that uses scores and references with a pre-defined temporal resolution and counts the number of misclassified segments. Such point-based measurement overly relies on this resolution and may not accurately measure misclassified ranges. To properly measure misclassified ranges and better evaluate spoof localization performance, we upgrade point-based EER to range-based EER. Then, we adapt the binary search algorithm for calculating range-based EER and compare it with the classical point-based EER. Our analyses suggest utilizing either range-based EER, or point-based EER with a proper temporal resolution can fairly and properly evaluate the performance of spoof localization.

Index Terms: partial spoof, metric, spoof localization, equal error rate, range-based

1. Introduction

Automatic speaker verification (ASV) is vulnerable to spoofing attacks (also known as presentation attacks or PA) [1]. Some challenges were thus held to encourage the development of countermeasures (CMs) to protect ASV from spoofing, such as ASVspoof [2–7] and ADD [8]. CMs for those challenges operate at the utterance level to detect whether an utterance is spoofed. The equal error rate (EER) and tandem detection cost function (t-DCF) [9] are then commonly used to evaluate the performance of CMs and consistently measure the progress in this field over time.

Partial Spoof (PS) [10] is a recently proposed spoofing scenario in which only a fraction of speech utterances are spoofed. It is one of the most important and challenging scenarios for the anti-spoofing community as detecting a fraction of speech segments is much more difficult than detecting a whole spoofed utterance. Accordingly, besides conventional utterance-level detection, spoof localization, also known as segment-level detection [11–13], was designed for the PS scenario. Spoof localization aims to locate spoofed regions within partially spoofed audio, that is, to answer *"when do spoofs happen?"*. Spoof localization is an important task for the PS scenario that can be used as a pre-processing step and provides cues to further analyze attackers' intentions.

It is also crucial to evaluate different models for spoof localization to make progress. However, as a newly introduced task in the anti-spoofing community, there is currently no established way of properly measuring the performance of spoof localization. Most metrics usually face dilemmas [14] and depend on a pre-defined threshold. Furthermore, the use of different measurements, such as counting the number of misclassified segments with fixed temporal resolutions (10 ms [15], 20 ms [10]) or measuring the duration of misclassified regions [13], hinders the comparison of different spoof localization methods across the literature. Following [16], we named the former approach of counting the number of misclassified segments "point-based" measurement and the latter approach of measuring the duration of misclassified regions "range-based" measurement.

For example, Yi *et al.* [13] used range-based precision, recall, and F1 to measure performance in accurately detecting spoofed regions. However, these require a pre-defined threshold and have a high bias on imbalanced data [17]. Furthermore, Zhang *et al.* [15] utilized point-based IoU (intersection over union, also known as the Jaccard index) by counting the number of accurately predicted frames to describe the similarity between reference and prediction. All of the above metrics depend on a pre-defined threshold. In contrast, we [12] adapted a threshold-free EER from the utterance level to the segment level. However, it is still a point-based EER that requires a pre-defined temporal resolution for reference and it is easy to ignore some misclassified regions that can only be measured at high precision. Range-based EER is a possible solution to measuring such misclassified regions properly.

To estimate range-based EER for measuring the performance of spoof localization in the PS scenario, we adapted the binary search algorithm (also known as the half-interval search method [18]). Then, we compared range-based EER with point-based EER for better understanding.

Our results show that when the temporal resolution of a point-based reference is coarser than the temporal resolution of the training data, point-based evaluation becomes too coarse. For fair and proper evaluation of spoof localization models, we recommend using range-based EER, or point-based EER that uses references with a finer temporal resolution than that used to train spoof localization models.

The remainder of this paper is structured as follows. Section 2 introduces the basic properties of point-based measurement and its relationship with range-based measurement. Section 3 describes range-based EER and proposes the adapted binary search algorithm to estimate range-based EER for spoof localization. Section 4 introduces the experiments and discusses the relationship between point-based and range-based EER. Finally, Section 5 gives the conclusion.

2. Point-Based Measurement

In this section, we first introduce the widely used point-based measurement. Then, we extend point-based measurement to

10.21437/Interspeech.2023-1214

Table 1: *Confusion matrix for spoof localization.*

		Hypothesis	
		Positive (*spoof*)	**Negative** (*bona fide*)
Ref.	**Positive** ($\mathcal{P}$)	TP	FN
	Negative ($\mathcal{N}$)	FP	TN

range-based measurement, providing a foundation for understanding the basic properties of range-based EER, which we will discuss in the next section.

2.1. Two types of errors: false positive and false negative

Following the ISO/IEC standard [1], we treat *spoof* as positive and *bona fide* as negative. Then, CMs are subject to two types of errors: false positive and false negative[1]:

- FP (False Positive): the number or duration of *bona fide* misclassified as *spoof*.
- FN (False Negative): the number or duration of *spoof* misclassified as *bona fide*.

The normalized (proportional) versions of FP and FN are called the false positive rate (FPR) and false negative rate (FNR). The corresponding confusion matrix is shown in Table 1, where TP (true positive) and TN (true negative) refer to correctly predicted *spoof* and *bona fide*, respectively. They can be calculated by either counting the number of segments or measuring the duration of eligible regions.

2.2. Classical point-based EER

Point-based EER is widely utilized for the binary classification task. It is a threshold-free metric and is the error rate with a specific threshold where the FPR is closest to the FNR. Following the predicted scores and confusion matrix defined in Table 1, we express the definition of FPR and FNR as follows:

$$P_{\text{FP}}(\tau) = \frac{1}{|\Lambda^p_{\mathcal{N}}|} \sum_{m \in \Lambda^p_{\mathcal{N}}} \mathbb{1}(s_m < \tau), \tag{1}$$

$$P_{\text{FN}}(\tau) = \frac{1}{|\Lambda^p_{\mathcal{P}}|} \sum_{m \in \Lambda^p_{\mathcal{P}}} \mathbb{1}(s_m \geq \tau), \tag{2}$$

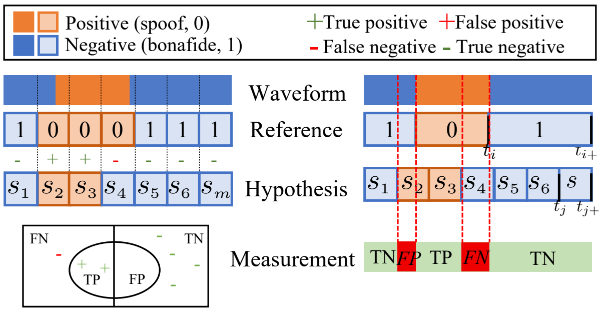
where both $P_{\text{FP}}(\tau)$ and $P_{\text{FN}}(\tau)$ are functions of a pre-defined threshold τ. $\Lambda^p_{\mathcal{N}}$ and $\Lambda^p_{\mathcal{P}}$ index bona fide and spoof *segments*, respectively. Then, $|\Lambda^p_{\mathcal{N}}|$ and $|\Lambda^p_{\mathcal{P}}|$ denote the total number of bona fide and spoof segments separately. s_m is the segment score for the m-th segment as shown in Fig. 1(a). $\mathbb{1}(\cdot)$ denotes the indicator function that outputs 1 when the condition is true and 0 otherwise.

EER is decided by $\hat{\tau}$ where the value of $P_{\text{FP}}(\hat{\tau})$ is infinitesimally close to $P_{\text{FN}}(\hat{\tau})$. Then, EER can be computed by:

$$EER = \frac{P_{\text{FP}}(\hat{\tau}) + P_{\text{FN}}(\hat{\tau})}{2}, \tag{3}$$

where

$$\hat{\tau} = \arg\min_{\tau} |P_{\text{FP}}(\tau) - P_{\text{FN}}(\tau)|. \tag{4}$$



(a) Point-based

(b) Range-based

Figure 1: *Comparison of point-based and range-based measurement. (In the Reference row of this example, there are 7 uniform points in (a) and 3 varied ranges in (b))*

2.3. From point-based to range-based measurement

Although the common evaluation method is to utilize point-based measurement, it can be implemented by range-based measurement for time series problems. Depending on the method of measuring the predicted results, all properties of the previously defined confusion matrix can be calculated using either the point-based or range-based[2] measurements as illustrated in Fig. 1.

As for the point-based measurement shown in Fig. 1 (a), we need to split the audio into uniform segments with a fixed resolution and assign corresponding reference labels. Then, we measure the performance on the basis of a comparison between those pre-segmented labels and discrete predicted segment scores. However, this point-based measurement can be easily influenced by the resolution of the references. This is because each uniform segment corresponds to only one label, and the segment is more likely to contain different classes with coarser resolution. This could result in imprecise evaluation results. Besides, a finer resolution can ensure that the segment is more likely to have only one class, thereby improving the precision of the evaluation.

In contrast, we do not need to do pre-segmentation for the range-based measurement as shown in Fig. 1 (b). Instead, we need to measure the duration of misclassified regions between references and hypotheses of each trial with higher precision. Thus, range-based measurement needs to record the boundaries for bona fide or spoof regions in the references and hypotheses. But it does not require a definition of resolution.

3. Range-Based Measurement

3.1. Range-based FPR, FNR, and EER

Although the implementation of measurements in existing literature may be different[3], all existing metrics are defined on the basis of the confusion matrix in Table 1 and can be calculated by using point-based or range-based measurement. Suppose a hypothesis is given by a segment-level detection model and each segment in the hypothesis has a score as shown in Fig. 1(b). Then, the range-based version of Eqs. (1) and (2) can be formu-

[1] Note that FA and MD in [2], and FR and FA in [19] are equivalent to FP and FN in this paper, respectively. Their names are different, but they share the same definition for the spoof scenario.

[2] Those two levels are called "classical" and "durative" in [20], and "frame-based" as well as "boundary-based" in [21].

[3] Different measurements are utilized in the PS scenario. Point-based measurement: EER in [12] and IoU in [15]. Range-based measurement: precision, recall, and F1 in [13].

lated as follows:

$$P_{\text{FP}}(\tau) = \frac{1}{D_{\mathcal{N}}} \sum_{i \in \Lambda^r_{\mathcal{N}}} \sum_{j} \mathbb{1}(s_j < \tau)\mathbb{T}(r_i, r_j), \tag{5}$$

$$P_{\text{FN}}(\tau) = \frac{1}{D_{\mathcal{P}}} \sum_{i \in \Lambda^r_{\mathcal{P}}} \sum_{j} \mathbb{1}(s_j \geq \tau)\mathbb{T}(r_i, r_j), \tag{6}$$

where i and j index the time range in the reference and hypothesis, separately. $\Lambda^r_{\mathcal{N}}$ and $\Lambda^r_{\mathcal{P}}$ index bona fide and spoof *ranges* in references. s_j is the predicted score derived from the CM for the j-th range[4]. $D_{\mathcal{N}}$ and $D_{\mathcal{P}}$ respectively denote the total duration of bona fide and spoof *ranges*, where $D_{\mathcal{N}} = \sum_{i \in \Lambda^r_{\mathcal{N}}} \mathbb{T}(r_i, r_i)$ and $D_{\mathcal{P}} = \sum_{i \in \Lambda^r_{\mathcal{P}}} \mathbb{T}(r_i, r_i)$. $\mathbb{T}(r_i, r_j)$ denote the overlapped duration between two ranges r_i and r_j. r_i is the range with the start time t_i and end time t_{i+1}. Then, $\mathbb{T}(r_i, r_j)$ can be formulated as:

$$\mathbb{T}(r_i, r_j) = \max(0, \min(t_{i+1}, t_{j+1}) - \max(t_i, t_j)). \tag{7}$$

However, calculating EER usually requires comparing $P_{\text{FP}}(\tau)$ and $P_{\text{FN}}(\tau)$ on the basis of all possible τ. Thus, we adapted the binary search algorithm to find $\hat{\tau}$ and estimate range-based EER.

3.2. Binary search algorithm for range-based EER

The binary search algorithm is a method that efficiently searches for a value in a sorted list of elements. The algorithm works by dividing the list in half at each iteration and determining which half of the list the target value is in. We adapted the binary search algorithm to estimate range-based EER as shown in Algorithm 1[5]. The notations used are shown in Table 2. Subscripts l, m, and r represent the left, middle, and right values respectively in the region of each iteration during binary search. To better adapt the binary search algorithm and estimate range-based EER, we made two modifications:

1. We divided the list in half on the basis of the quantile but not the value as shown in line 13 of Algorithm 1. Given that the EER is calculated by score distribution, we searched $\hat{\tau}$ on the basis of the *quantile* ($\frac{Q_l+Q_r}{2}$) but not the value ($\frac{\tau_l+\tau_r}{2}$) as in the original binary search algorithm.
2. We introduced an additional condition $(P_{\text{FP}}(\tau_l) - P_{\text{FN}}(\tau_l)) \times (P_{\text{FP}}(\tau_m) - P_{\text{FN}}(\tau_m)) \leq 0$ to assign the value for the middle threshold as shown in line 8. When we have a predicted score that refers to how likely it would be bona fide, $P_{\text{FP}}(\tau)$ is an increasing function while $P_{\text{FN}}(\tau)$ is a decreasing function of the threshold τ. Here is an important theorem, that is, when $\tau < \hat{\tau}$, $P_{\text{FP}}(\tau) < P_{\text{FN}}(\tau)$ and vice versa [22]. Thus, in addition to the common condition in the binary search algorithm, we utilized $(P_{\text{FP}}(\tau_l) - P_{\text{FN}}(\tau_l)) \times (P_{\text{FP}}(\tau_m) - P_{\text{FN}}(\tau_m)) \leq 0$ to assign the value for the middle threshold τ_m.

4. Experiments

To further explore the relationship between point-based and range-based measurement, we measured EER on a recent powerful model [10] for spoof localization. This section introduces the database and experimental configuration.

Table 2: *Notations used in the binary search algorithm.*

s	List of sorted predicted scores,
y	List of ground-truth labels,
τ, Q_τ	Threshold and its quantile, $Q_\tau \in [0, 100]$,
τ_l, Q_l	Lower threshold and its quantile,
τ_r, Q_r	Upper threshold and its quantile,
τ_m, Q_m	Middle threshold and its quantile,
$prec$	Precision we want to get.

Algorithm 1 Binary search algorithm for range-based EER.

```
Input  : $s, y, prec$
Output: Estimated range-based $EER$
1  Function Cal_FPR_FNR($s$, $y$, $\tau$):
2      //Utilize Eqs. (5) and (6) to
       calculate range-based FPR and FNR on
       the basis of the threshold $\tau$.
3  return $P_{FP}(\tau)$, $P_{FN}(\tau)$
4  Function Percentile($s$, $Q$):
5      //Get percentile value of $Q$ in $s$.
6  return percentile value
7  while $\tau_l \leq \tau_r$ AND $abs(P_{FP}(\tau_m) - P_{FN}(\tau_m)) \geq prec$ do
8      if $(P_{FP}(\tau_l) - P_{FN}(\tau_l)) \times (P_{FP}(\tau_m) - P_{FN}(\tau_m)) \leq 0$ then
9          //when $\tau_l < \hat{\tau} < \tau_m$
           $\tau_r \leftarrow \tau_m, Q_r \leftarrow Q_m$
           $P_{\text{FP}}(\tau_r) \leftarrow P_{\text{FP}}(\tau_m), P_{\text{FN}}(\tau_r) \leftarrow P_{\text{FN}}(\tau_m)$
10     else
11         //when $\tau_m < \hat{\tau} < \tau_r$
           $\tau_l \leftarrow \tau_m, Q_l \leftarrow Q_m,$
           $P_{\text{FP}}(\tau_l) \leftarrow P_{\text{FP}}(\tau_m), P_{\text{FN}}(\tau_l) \leftarrow P_{\text{FN}}(\tau_m)$
12     end
13     $Q_m \leftarrow \lfloor \frac{Q_l+Q_r}{2} \rfloor$
       $\tau_m \leftarrow$ Percentile($Q_m$)
       $P_{\text{FP}}(\tau_m), P_{\text{FN}}(\tau_m) =$ Cal_FPR_FNR($s, y, \tau_m$)
14 end
15 $EER = \frac{P_{\text{FP}}(\tau_m)+P_{\text{FN}}(\tau_m)}{2}$
   return $EER$
```

4.1. Database

We used the publicly available PartialSpoof[6] database to calculate EER. The PartialSpoof database [23] was generated by randomly substituting spoof (or bona fide) speech segments as bona fide (or spoof) from the same speaker. Bona fide and spoof segments were concatenated using the overlap-add method.

4.2. Configuration

We measured the EER on the most powerful CM [10] on the PartialSpoof database when we wrote this paper. It utilized the self-supervised learning (SSL) model w2v2-large [24] as the front-end, gMLP [25] as the back-end, P2SGrad-based mean squared error [26] as the loss function, and Adam as the optimizer. It supports training using multiple resolutions or a single resolution. The finest resolution is at a frame level of 20 ms based on the configuration of the SSL model, and the coarsest segment-level resolution is 640 ms. We compared outputs extracted from branches of different resolutions trained at multiple resolutions and then discussed the relationship between them. We set $prec = 1e-5$ to estimate the range-based EER.

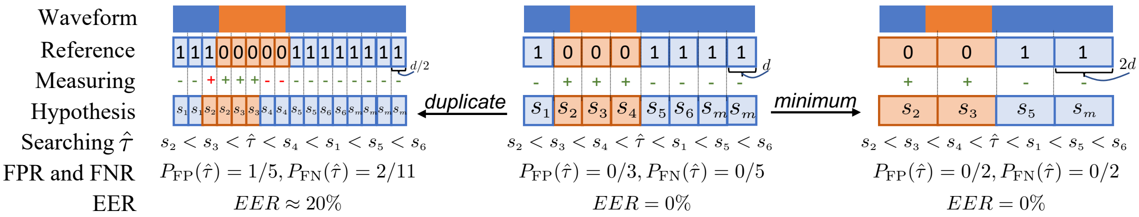
[4] The duration of the j-th range in the hypothesis can either be uniform with a resolution of $d = t_{j+1} - t_j$ or a variable length of $t_{j+1} - t_j$. Experiments in Section 4 of this paper belong to the former case.

[5] https://github.com/nii-yamagishilab/PartialSpoof

[6] https://zenodo.org/record/5766198

Table 3: *Range-based and point-based EER (%) of multi-reso. CM in PartialSpoof.*

	Development set								Evaluation set							
Reso. of	Range-	Point-based EER							Range-	Point-based EER						
Training	based EER	10	20	40	80	160	320	640	based EER	10	20	40	80	160	320	640
20	24.39	23.48	*0.86*	0.84	0.83	0.82	0.82	0.77	30.40	29.78	*12.84*	11.94	10.52	8.42	5.96	4.06
40	24.48	23.61	23.62	*0.83*	0.81	0.80	0.79	0.74	30.11	29.93	29.94	*11.94*	10.51	8.43	5.98	4.10
80	24.60	23.56	23.56	23.56	*0.81*	0.79	0.78	0.71	30.65	30.12	30.12	30.15	*10.92*	8.70	6.14	4.15
160	25.55	24.37	24.37	24.36	24.39	*0.79*	0.77	0.72	31.36	30.49	30.50	30.52	30.56	*9.24*	6.40	4.11
320	30.03	28.99	28.99	28.99	29.02	29.09	*0.75*	0.69	33.91	33.39	33.38	33.41	33.45	33.48	*6.34*	3.97
640	34.96	34.84	34.84	34.85	34.87	34.87	34.59	*2.15*	37.38	37.53	37.53	37.54	37.56	37.56	37.54	*5.19*


Waveform
Reference
Measuring
Hypothesis
Searching $\hat{\tau}$
FPR and FNR
EER
1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1
$d/2$
duplicate
$s_2 < s_3 < \hat{\tau} < s_4 < s_1 < s_5 < s_6$
$P_{\mathrm{FP}}(\hat{\tau}) = 1/5, P_{\mathrm{FN}}(\hat{\tau}) = 2/11$
$EER \approx 20\%$
1 0 0 0 1 1 1 1
d
s_1 s_2 s_3 s_4 s_5 s_6 s_m s_m
$s_2 < s_3 < s_4 < \hat{\tau} < s_1 < s_5 < s_6$
$P_{\mathrm{FP}}(\hat{\tau}) = 0/3, P_{\mathrm{FN}}(\hat{\tau}) = 0/5$
$EER = 0\%$
minimum
0 0 1 1
$2d$
s_2 s_3 s_5 s_m
$s_2 < s_3 < s_4 < \hat{\tau} < s_1 < s_5 < s_6$
$P_{\mathrm{FP}}(\hat{\tau}) = 0/2, P_{\mathrm{FN}}(\hat{\tau}) = 0/2$
$EER = 0\%$

(a) Fine-grained resolution ($d/2$) *(b) Source resolution (d)* *(c) Coarser-grained resolution (2d)*

Figure 2: *An example for changing of score and error rate when up-sampling predicted score to fine-grained level (left) and down-sampling to coarser-grained level (right). [Given a threshold $\hat{\tau}$, the symbols "+" and "-" represent the predicted class as positive (spoof) and negative (bona fide) respectively. The color green indicates a correct prediction, while red represents a false prediction.]*

4.3. Up-sampling and down-sampling predicted scores

Models trained at fixed resolutions can usually only produce scores for uniform segments of the same resolution used during training. Then, we usually evaluate the performance using pre-segmented labels with the same resolution. If we want to measure the performance at different resolutions, we need to perform additional post-processing, like up-sampling or down-sampling, to convert predicted scores to the target measurement resolution. In this paper, as shown in the Hypothesis row of Fig. 2, (1) when up-sampling predicted scores to a fine-grained resolution, we duplicated each segment score following the relationship between the source resolution and fine-grained resolution, and (2) when down-sampling scores to a coarser-grained resolution, we aggregated adjacent segments by selecting their minimum[7] value.

4.4. Results and discussion

Table 3 shows the results for models evaluated on the development and evaluation set of PartialSpoof separately. Each row has the same original predicted scores, and each column presents the measurement resolution. Thus, the point-based EER on the diagonal presents the training and measuring at the same resolution. The upper triangle of tables, from left to right, shows down-sampling predicted scores to coarser-grained resolution, and the lower triangle of those tables, from right to left, shows up-sampling predicted scores to fine-grained resolution.

From the point-based EER in Table 3, we can notice that although each row has the same predicted score, different measurement resolutions can lead to significantly different performances. This is because the point-based references were defined from the pre-defined resolution. In the upper triangle (i.e., where the temporal resolution of the point-based reference is coarser than the temporal resolution of the training data), point-based EER may be an "underestimation" in terms of the spoof localization performance, since the reference becomes too coarse and does not reflect accurate boundary information as shown in Fig. 2. In such cases, although the error value becomes smaller, it just indicates that the task is easier and does not mean that spoof localization is more accurate. In general, errors must be interpreted with caution and in consideration of the temporal resolution used for the point-based EER. On the other hand, when the temporal resolution of the point-based reference is finer than the temporal resolution of the training data (lower triangle of Table 3), or a range-based reference is used (column "Range-based EER" of Table 3), the error value will be naturally larger since the reference is more accurate and we can therefore account for errors at a finer level. Note that for the same row, even if the error is higher, it does not mean that the model is inaccurate – they shared the same original predicted scores but were evaluated on references with different temporal resolutions.

Thus, for fair and proper evaluation of spoof localization models, we recommend using range-based EER, or point-based EER that uses references with a finer temporal resolution than that used during model training. In addition, when the training temporal resolution is unknown, the range-based EER would be a more appropriate choice.

5. Conclusion

In this paper, we first defined range-based EER for spoof localization and then adapted the binary search algorithm to estimate it. We finally utilized range-based EER and classical point-based EER to analyze the performance of spoof localization deeply and discussed the relationship between them. For the measurement, we recommend using range-based EER, or point-based EER with unseen and finer temporal resolutions compared with the training resolution to more fairly and properly evaluate the performance of spoof localization.

6. Acknowledgements

This study is partially supported by the Japanese-French joint national VoicePersonae project supported by JST CREST (JPMJCR18A6, JPMJCR20D3), JPMJFS2136 and the ANR (ANR-18-JSTS-0001), MEXT KAKENHI Grants (21K17775, 21H04906, 21K11951), and Google AI for Japan program.

[7] A segment with a lower score is more likely to be spoofed.

7. References

[1] I. J. S. Biometrics, "Iso/iec 30107: Information technology — biometric presentation attack detection," 2016.

[2] 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 *Proc. Interspeech*, 2015, pp. 2037–2041.

[3] 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 *Proc. Interspeech*, 2017, pp. 2–6.

[4] 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.

[5] 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.

[6] 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 2021 Workshop*, 2021, pp. 47–54.

[7] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint arXiv:2210.02437*, 2022.

[8] 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.

[9] T. Kinnunen, K. A. Lee, N. Evans, M. Todisco, J. Yamagishi, and D. A. Reynolds, "t-DCF : a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification," in *Proc. Odyssey 2018*, 2018.

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

[12] 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. [Online]. Available: http://arxiv.org/abs/2104.02518

[13] 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.

[14] M. Gösgens, A. Zhiyanov, A. Tikhonov, and L. Prokhorenkova, "Good classification measures and how to find them," in *Proc. NeurIPS 2021*, vol. 34, 2021, pp. 17 136–17 147.

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

[16] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," in *Proc. NeurIPS 2018*, 2018, p. 1924–1934.

[17] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.

[18] L. F. Williams, "A modification to the half-interval search (binary search) method," in *Proc. Annual Southeast Regional Conference*. New York, NY, USA: Association for Computing Machinery, 1976, p. 95–101. [Online]. Available: https://doi.org/10.1145/503561.503582

[19] X. Wang and J. Yamagishi, "A practical guide to logical access voice presentation attack detection," in *Frontiers in Fake Media Generation and Detection*. Springer, 2022, pp. 169–214.

[20] S. M. Modaresi, A. Osmani, M. Razzazi, and A. Chibani, "Uniform evaluation of properties in activity recognition," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2022, pp. 83–95.

[21] S. Tong, N. Chen, Y. Qian, and K. Yu, "Evaluating vad for automatic speech recognition," in *2014 12th International Conference on Signal Processing (ICSP)*. IEEE, 2014, pp. 2308–2314.

[22] N. Poh and S. Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, 2006.

[23] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, "PartialSpoof Database - Partially Spoofed Audio Dataset for Anti-spoofing," May 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5766198

[24] 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.

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

[26] X. Wang and J. Yamagishi, "A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection," in *Proc. Interspeech*, 2021, pp. 4259–4263.