J-SPAW: Japanese speaker verification and spoofing attacks recorded in-the-wild dataset

Sayaka Shiota[1], *Suzuka Horie*[1], *Kouta Kanno*[1], *Shinnosuke Takamichi*[2,3]

[1]Tokyo Metropolitan University, Japan
[2]Keio University, Japan
[3]The University of Tokyo, Japan

sayaka@tmu.ac.jp,horie-suzuka@ed.tmu.ac.jp,kanno-kouta@ed.tmu.ac.jp,
shinnosuke_takamichi@keio.jp,

Abstract

In this paper, we present J-SPAW (Japanese speaker verification and spoofing attacks recorded in-the-wild dataset), a novel speech database designed for speaker verification and spoofing detection in-the-wild environments[1].J-SPAW is a unique database that simultaneously evaluates speaker verification and spoofing detection under realistic conditions, focusing on physical access scenarios, including replay attacks. The database includes diverse physical access scenarios, enhancing the variety and applicability of datasets available for anti-spoofing research. Our experimental results demonstrate that J-SPAW enables comprehensive analysis of spoofing detection from various perspectives and can be utilized for both speaker verification and spoofing detection tasks. This contribution is expected to advance state-of-the-art speaker verification and spoofing detection and provide a valuable resource for future research.

Index Terms: Automatic speaker verification, spoofing attacks, replay attack, speech dataset, in-the-wild recording

1. Introduction

In recent years, the spreading of smartphones and the expansion of voice assistant usage have heightened the importance of automatic speaker verification (ASV) [1]. ASV is a technology that identifies individuals using unique information in their voices, and it holds great promise for various applications. However, as ASV technology advances, the threat of spoofing attacks has also increased, making it urgent to develop countermeasures to ensure the reliability of ASV systems [2]. Spoofing attacks can be broadly classified into physical access (PA) attacks and logical access (LA) attacks, with active research driven by contributions from the ASVspoof challenges [3, 4].

For PA attacks, the collection of replay attack data is necessary, and there are few databases outside of competitions that can simultaneously evaluate ASV and spoofing detection. Notable PA databases include ASVspoof 2017 [5] and ASVspoof 2019 [6]. ASVspoof 2017 provides data that considers various replay attack scenarios, including the impact of recording environments and playback devices. ASVspoof 2019 offers a larger-scale database considering more diverse environments and devices, allowing for integrated evaluation with ASV. However, these databases are primarily designed for competitions and lack sufficient discussion on speech data collected in the wild environments. PA-specific databases such as ReMASC [7] and AVspoof [8] exist. ReMASC considers more diverse recording environments and is particularly suitable for evaluating the impact of replay attacks in mobile device environments. AVspoof covers various replay attack scenarios, considering various source recording and playback devices. However, these databases do not aim to simultaneously evaluate both ASV and anti-spoofing, making consistent evaluation challenging.

As such, existing PA databases each have their characteristics, but there are limited datasets that can evaluate the impact of diverse replay attack environments while simultaneously considering speaker verification. To address this, we have constructed a new database, J-SpAW (Japanese speaker verification and spoofing attacks recorded in-the-wild dataset), for Japanese ASV and spoofing detection. This database considers both ASV and spoofing detection scenarios. It is recorded in environments that simulate situations where attackers might realistically obtain voice samples without the consent of an authorized speaker. Specifically, it includes diverse speech content, speaker attributes, recording environments, and various types of spoofed speech.

Our experiments demonstrate that J-SpAW performs well as an evaluation set for speaker verification. Furthermore, in spoofing detection, we confirmed that the recording environments for genuine speech and replay attacks are meticulously labeled, allowing for detailed analysis of their impacts. We also evaluated J-SpAW using the ASVspoof baseline model and one of the state-of-the-art model, reporting on the differences in characteristics and trends between the models.

2. Related works

2.1. Automatic speaker verification and spoofing countermeasure

ASV is a binary classification task that determines whether the input speech matches the registered speech of the same speaker. It involves comparing the features of the input speech sample with those stored in the system to verify the speaker's identity. As the accuracy of ASV improves, concerns about spoofing attacks have similarly increased. Various spoofing attacks are broadly classified into PA attacks using playback recordings and LA attacks using synthetic speeches. Various countermeasures (CM) have been proposed to detect these attacks. For PA attacks, many studies focus on the noise during nonconsensual recordings (source recordings) or replayed spoofed speeches and the differences between human liveness characteristics and loudspeaker playback characteristics [9, 10]. For LA attacks, numerous studies use various text-to-speech (TTS) synthesis and voice conversion (VC) techniques in training to model the unique characteristics of synthetic speeches [11, 12].

[1]The J-SPAW database is publicly available at: https://github.com/takamichi-lab/j-spaw/blob/main/README_en.md

Table 1: *Comparison of datasets in terms of tasks (spoof detection or ASV), language, source recording type, and environment of replayed recordings.*

Dataset	Spoof	ASV	Lang	Source recording	Replayed env.
AVspoof [8]	PA/LA		En	Controlled	Clean
ASVspoof2015 [3]	LA	✓	En	-	-
BTAS [13]	PA/LA		En	Controlled	Various
ASVspoof2017 [5]	PA	✓	En	Controlled	Various
ASVspoof2019 [6]	PA/LA	✓	En	Artificial	Clean
ASVspoof 2021 [14]	PA/LA	✓	En	Artificial	Clean
ReMASC [7]	PA		En	Realistic	Various
VSDC [15]	PA		En	Controlled	Various
LRPD [16]	PA		En	Artificial	Clean
J-SpAW (ours)	PA/(LA)	✓	Ja	Realistic	Various

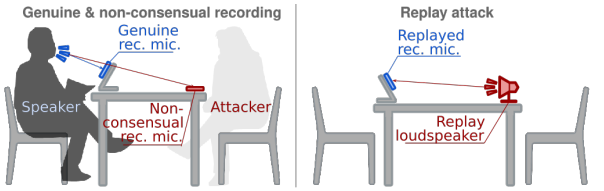

Figure 1: *Illustration of genuine recording, non-consensual recording (source recording), and replay attack. Voices by the non-consensual recording are used in the replay attack.*

2.2. Datasets for spoofing attack detection

Various datasets for spoofing attacks have been released; however, not many include PA attacks. Additionally, there are differences in whether the datasets target spoofing detection tasks alone or also include the ASV task, as well as in the recording environments of the spoofed speeches. These differences are summarized in Table 1. As shown in Table 1, few datasets include both spoofing and ASV tasks, and even fewer assume recording in realistic environments. Furthermore, while the source recordings in ReMASC are characterized by recording the genuine speaker's speeches close to the attacker's mouth, our study assumes non-consensual source recordings, where the genuine speaker's voice is recorded without their consent and from a distance from the recording microphone. This distinguishes our dataset from others.

3. Dataset construction strategy

The dataset consists of two categories of speech: genuine speech, which humans speak, and spoofed speech, which is generated through PA attacks conducted by an attacker using recorded playback. Each speech utterance is labeled as either genuine or spoofed and includes additional metadata such as speaker information, recording environment details during non-consensual recording or spoofing attacks, and other relevant conditions. This allows the dataset to be utilized for both anti-spoofing and ASV tasks. Figure 1 illustrates the recording methods for genuine recordings, non-consensual recordings (source recordings), and replay attacks, while Figure 2 shows the recording environments during the collection of spoofed speech.

3.1. Text material and participants recruit

We designed a text set consisting of 50 sentences in total: 25 voice command sentences that include a wake-up word such as "OK Google," and 25 dailylife-style sentences. Of these, 25

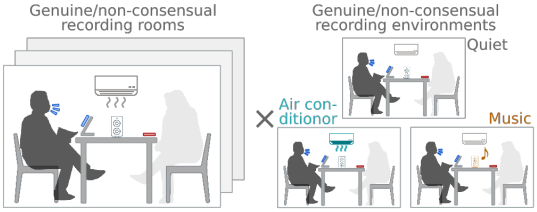

Figure 2: *Examples of recording conditions. Similarly, some conditions are used in replay attack.*

Table 2: *Recording and Replay conditions.*

Label	Description
Genuine recording microphone	
M1	Google Pixel 3 (1.0m from the speaker)
M2	Apple iPhone 8 (same position as above)
Non-consensual recording microphone	
M3	Apple iPad mini 5th generation
Genuine/non-consensual recording room	
R1	Room ($4.4(\mathrm{W}) \times 7.4(\mathrm{L}) \times 2.5(\mathrm{H})$ [m]) at institution 1
R2	Outside closing to a road at institution 1
R3	Room ($10.8(\mathrm{W}) \times 2.0(\mathrm{L}) \times 2.8(\mathrm{H})$ [m]) at institution 2
R4	Outside with lawn at institution 2
Recording environment	
E1	Quiet (R1, R3)
E2	Air conditioner is activating (R1, R3).
E3	Music is playing from a smart speaker (R1, R3).
E4	Outside (R2, R4)
Replay room	
r1	Room ($11.0(\mathrm{W}) \times 8.0(\mathrm{L}) \times 2.6(\mathrm{H})$ [m]) at institution 2
Replay loudspeaker	
s1	Bose Soundlink Micro Bluetooth Speaker Bundle
s2	iPad
s3	MacBook Pro
s4	Sony SRS-ZR7
Replay environment	
e1, e2, e3	Same to E1, E2, E3, respectively. e3 and E3 are different in songs used.
Replayed recording microphone	
m1, m2	Same to M1, M2 respectively.

voice command sentences and 20 daily-life style sentences are crafted to ensure a balanced distribution of phonemes in the target language (Japanese), while the remaining 5 daily-life style sentences are used as evaluation data.

We recruited participants (speakers) via snowball sampling. The participants were 21 male and 19 female native Japanese speakers. Each participant received 3000 JPY for their participation.

3.2. Genuine and non-consensual speech recording

As shown in Figure 1, both genuine and non-consensual speech recordings by an attacker are carried out simultaneously. The scenario assumes a situation where the attacker records the genuine speaker's speech without consent, with the non-consensual recording occurring at a distance from the genuine speaker. Genuine speech is treated as the actual speech in the context of spoof detection and as the speech of the registered speaker in ASV. Speech recorded through non-consensual recording can be considered source recording for performing a replay attack. The speaker sits in a chair and speaks toward a smartphone placed in front of them; the smartphone records the speech. This setup simulates the speaker speaking directly into their

smartphone. However, we fixed the smartphone onto a stand on a desk to avoid changes in acoustic characteristics if the smartphone were handheld.This smartphone is named the *genuine recording microphone*. Assuming the attacker is seated across from the speaker, a tablet is placed on the attacker's desk to record the genuine speaker's speech. This recorded speech is regarded as the *non-consensual utterance*, and we call the tablet the *non-consensual recording microphone*. We carry out these recordings in multiple locations and acoustic environments. We refer to the location as the *genuine/non-consensual recording room*[2] and the acoustic environment as the *genuine/non-consensual recording environment*. We simulate a typical daily acoustic environment for indoor recording, e.g., air-conditioner noise and background music. Under each condition, every participant (speaker) utters the text designed in Section 3.1. For all microphones, the sampling frequency, audio format, and number of channels are set to 48 kHz, RIFF WAV format, and single channel, respectively.

3.3. Physical access recording

Spoofed speech is recorded by replaying non-consensually recorded speech through a loudspeaker and capturing it with a smartphone. The room, the acoustic environment, the playback device, and the recording device used for this attack are referred to as the *Replay Room*, *Replay Environment*, *Replay Loudspeaker*, and *Replayed Recording Microphone*, respectively. The sampling frequency and format are the same as those described in Section 3.2. The speech captured by the replayed recording microphone is treated as a spoofed utterance in a spoofing attack.

3.4. Transfer function canceling

The performance of a replay attack depends on the replay environment, replay loudspeaker, and replayed recording microphone. Assuming a linear time-invariant acoustic characteristic, let x be the non-consensual utterance, h be the acoustic transfer function of the attack environment, and y be the observed signal. Then, $y = h * x$. If h can be canceled, it is expected that y would match x, making the replay attack more likely to succeed. Hence, we measure h by playing an impulse response measurement signal (time-stretched pulse) through the replay loudspeaker and recording it with the replayed recording microphone. We then convolve x with the inverse filter $h^{(\mathrm{inv})}$, obtaining $x' = h^{(\mathrm{inv})} * x$ as a new non-consensual utterance, and use x' to conduct the replay attack. In a realistic setting, it is not feasible for the attacker to directly access the recordings of the replayed recording microphone. However, it is possible to simulate this scenario. For example, one might estimate the impulse response based on captured images or perform a computer simulation that models the replay environment. Another simpler method would be for the attacker to place a new microphone near the replayed recording microphone to measure the impulse response. In this section, we regard the impulse response we measure as the ideal value that could be obtained by such methods.

4. Experiment

We evaluated the performance of J-SpAW using ASV and spoofing detection. The recording devices and environments used in this experiment are summarized in Table 2.

[2]Although we use the word "room," Table 2 shows that this may include outdoor environments.

Table 3: *EER (%) for ASV evaluation using J-SpAW and VoxCeleb1 evaluation data.*

Model	J-SpAW	Voxceleb1
ECAPA-TDNN[18]	1.75	0.86
ResNetSE34V2[19]	2.99	1.02
RawNet3[19]	1.87	0.89

4.1. Experimental condition

4.1.1. ASV evaluation

As described in Section 3.2, J-SpAW enables speaker verification evaluation using genuine speech recordings. Therefore, we assessed speaker verification performance using several state-of-the-art speaker verification models. All models were pre-trained on VoxCeleb2 [17], and the three models used in the evaluation were ECAPA-TDNN [18], ResNetSE34V2 [19], and RawNet3 [19]. Evaluation was conducted using the Equal Error Rate (EER).

The data composition for ASV evaluation in J-SpAW is as follows: A total of 8,000 genuine speech utterances (40 speakers × 50 utterances × 4 recording environments) were recorded, where the four recording environments correspond to E1–E4 in Table 2. Five utterances per speaker were selected for the ASV evaluation set, resulting in 800 utterances. Following the approach of VoxCeleb1, 7,600 genuine trials and 30,000 imposter trials were prepared, yielding a total of 37,600 trials. The genuine recording microphones used were a Pixel 3 (Pixel) and an iPhone 8 Plus (iPhone). During recording, both devices were placed side by side and recorded simultaneously. However, time synchronization was not performed.

4.1.2. Physical access evalution

We evaluate the spoofing detection performance of J-SpAW in the PA task. For the PA task, 25 voice command sentences were recorded through non-consensual recording and conducted simultaneously with genuine speech recording. A total of 4,000 utterances (40 speakers × 25 utterances × 4 recording environments) were utilized as source recordings for replay. The replayed recording microphones used the identical Pixel and iPhone as the genuine speech recording. The recording environments for the spoofed audio were silence, air conditioning, and music, which are referred to in e1 to e3, respectively, as shown in Table 2. The replayed recording microphones were performed using the same Pixel and iPhone as the genuine speech recording. The recording environments for the spoofed audio were silence, air conditioning, and music, where reffered in e1 to e3, respectively as shown in Table 2. We conducted spoofing detection experiments using 96,000 replay attacks (4,000 utterances x 3 replay environments x 4 replay loudspeakers x 2 replayed recording microphones) and 800 utterances (40 speakers x 5 utterances x 4 recording environments) that were the same as the ASV evaluation set in J-SpAW.

These utterances were evaluated using two spoofing detection models: the pre-trained Linear Frequency Cepstral Coefficients Gaussian Mixture Model (LFCC-GMM) [20, 21], which is published as a baseline system in ASVspoof2021, and wav2vec2.0 and AASIST (w2v2+AASIST) [22], which had one of the best performance in ASVspoof2021. The evaluation also used the EER, which is similar to ASV; however, its meaning differs. It indicates the point at which the genuine speech rejection rate equals the spoofed speech acceptance rate. A higher

Table 4: *EER(%) for spoofing detection (LFCC-GMM/w2v2.0+AASIST)*

Conditions	r1	m1	m2	e1	e2	e3	s1	s2	s3	s4	Pooled
R1	46.94/4.13	48.20/2.87	46.35/5.35	42.86/4.13	50.16/3.80	48.87/4.75	60.08/1.59	21.24/7.04	40.07/5.40	57.78/1.29	46.94/4.13
R2	34.27/0.95	36.02/0.73	33.33/0.97	29.40/0.95	37.96/0.95	34.44/0.95	41.10/0.02	16.03/1.68	25.94/0.95	43.81/0.00	34.27/**0.95**
R3	54.04/1.06	51.66/0.43	55.87/1.40	44.59/0.72	59.59/0.39	56.90/1.75	73.33/0.31	20.35/1.11	47.40/1.82	70.18/0.09	54.04/**1.06**
R4	54.61/0.81	54.74/0.11	53.54/0.99	47.45/0.89	59.87/0.17	54.74/0.84	78.95/0.02	27.37/0.96	43.16/1.12	71.74/0.00	54.61/**0.81**
E1	44.51/2.00	44.09/1.42	44.99/2.50	38.07/2.00	48.50/2.00	46.98/2.50	59.50/0.50	18.07/3.11	37.88/2.50	57.53/0.43	44.51/2.00
E2	45.11/2.48	44.98/1.51	45.50/3.03	38.95/2.10	48.88/2.06	47.56/2.95	60.00/0.48	19.50/3.50	38.55/3.46	58.50/0.42	45.11/2.48
E3	54.50/4.38	54.50/3.43	55.00/5.38	48.54/4.00	58.00/4.00	57.00/4.50	68.92/1.50	28.62/6.00	49.54/5.50	68.00/1.07	54.50/**4.38**
E4	45.00/1.50	44.55/1.00	45.00/2.01	38.98/1.50	48.02/1.50	46.58/1.95	60.00/0.40	19.09/2.90	38.58/2.06	57.44/0.12	45.00/1.50
Pooled	46.88/2.64	46.50/1.88	47.12/3.49	41.00/2.62	50.50/2.60	49.38/3.00	62.11/**0.74**	**21.77/4.24**	40.89/3.74	59.75/**0.62**	46.88/2.64

EER indicates that spoofing detection failed. Since the spoofing attack was successful, it can be said that it is more challenging to detect spoofing attacks.

4.2. Experimental results

4.2.1. ASV evaluation

Table 3 presents the EERs of ASV for each model. For all three models, the EER of J-SpAW is slightly higher than that of VoxCeleb1. It can be considered that these pre-trained models were trained on VoxCeleb2, which differs in language and domain from J-SpAW. However, the obtained EER values are sufficiently low, indicating that J-SpAW is not a particularly difficult task for ASV and is adequate as an ASV system to be used after spoofing detection.

4.2.2. Physical access evaluation

Table 4 presents the EERs for spoofing detection for both the LFCC-GMM and AASIST models. The table rows delineate the EER for each environment during source recordings, whereas the columns illustrate the EER for each environment during replayed recordings. The overall EER, as shown in the bottom right corner of the table, was 46.88% for LFCC-GMM and 2.64% for w2v2+AASIST. Since LFCC-GMM is not inherently a high-performance detection model, it was expected that the detection performance for J-SpAW would not be very high. On the other hand, the EER for w2v2+AASIST is low, indicating successful spoofing detection. Comparing the rows in the Pooled section, LFCC-GMM shows consistently high EER across all recording environments. Except for the case where replay was performed using s2 (iPad), it failed to detect spoofing effectively. Similarly, when comparing the Pooled columns, EER remains above 34% in all conditions. On the other hand, w2v2+AASIST achieves significantly higher spoofing detection accuracy, making the impact of recording conditions more apparent compared to LFCC-GMM. The EER is below 1% in replay recording when using s1 (Bose) and s4 (Sony) loudspeakers. However, it is higher when using s2 (iPad). This suggests that while w2v2+AASIST successfully learns the playback characteristics of high-end speakers, its detection performance may degrade for lower-quality speakers. Regarding the recording conditions for genuine speech, R2, R3, and R4 exhibit high spoofing detection performance, whereas R1 and E3 show slightly higher EER. In particular, E3 represents a non-consensual recording scenario with background music, making spoofing detection more challenging.

When additional combinations were compared, the combination of e2 (air-conditioned) and s4 (Sony) achieved the highest detection performance with w2v2+AASIST. Table 5 presents the results after applying the transfer function canceling method (described in Section 3.4) to mitigate the impact of the loudspeaker and room transfer function during replay attacks. Consequently, the EER increased slightly, suggesting that reducing the influence of the replay attack environment makes spoofing detection more challenging.

Table 5: *EER (%) of spoofing detection without and with transfer function canceling*

	w/o canceling	w/ canceling
w2v2+AASIST	0.52	1.12

Table 6: *Uncorrected p-values between factors by ANOVA. The smaller this value is, the more there is an interaction between the two factors that affect EER.* ***Bold*** *indicates* $p < 0.05$.

Factor 1	Factor 2	LFCC-GMM	w2v2+AASIST
R1–R4	e1–e3	0.995	0.907
R1–R4	m1–m2	0.894	0.102
R1–R4	s1–s4	**0.000**	**0.000**
E1–E4	e1–e3	1.000	1.000
E1–E4	m1–m2	1.000	0.893
E1–E4	s1–s4	1.000	**0.003**
e1–e3	m1–m2	**0.026**	0.408
e1–e3	s1–s4	**0.000**	0.242
m1–m2	s1–s4	0.293	**0.030**

4.3. Impact of recording / attack conditions

Finally, we investigate the interaction between the source and replay attack recording conditions. Because each condition can be set independently, one would intuitively expect that each condition affects the EER independently and that no entanglement occurs between conditions. To verify this, we used ANOVA to compute the two-way interactions between conditions. Table 6 shows the p-values for these interactions. In most of the condition pairs, there is no interaction. However, for both models, there is a strong interaction between the recording room and the replay loudspeaker. The EER appears to be strongly influenced by combining these two factors. Moreover, for LFCC-GMM and w2v2+AASIST, different factors exhibit interactions. Model traits emerge through interactions. Our corpus is expected to contribute to evaluating such interactions and training models with minimal inter-condition interaction.

5. Conclusion

Replay attacks, a type of physical access attack, require source and replay recordings, limiting the diversity of existing databases. To enhance diversity and enable analysis of recording environments and their relationship with ASV, we constructed a new J-SpAW database and evaluated its performance. Future challenges include expanding the variety of spoofing recording environments, addressing logical access attacks, and developing speech datasets that further increase the difficulty of spoofing detection.

6. References

[1] Z. Rui and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *Transaction on IEEE Access*, vol. 7, pp. 5994–6009, 2019.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Transaction on Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "Asvspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 15, p. 3750, 2014.

[4] https://www.asvspoof.org/.

[5] 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 proceedings of Interspeech*, pp. 2–6, 2017.

[6] 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," *Transactions on IEEE Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[7] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "Remasc: Realistic replay attack corpus for voice controlled systems," *In proceedings of Interspeech*, 2019.

[8] S. K. Ergüay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," *In proceedings of IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6, 2015.

[9] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures," *Transaction on Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 513–566, 2023.

[10] R. Yaguchi, S. Shiota, N. Ono, and H. Kiya, "Replay attack detection based on spatial and spectral features of stereo signal," *Transaction on Journal of Information Processing*, vol. 29, pp. 275–282, 2021.

[11] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e2, 2020.

[12] N. M. Müller, F. Dieckmann, P. Czempin, R. Canals, and K. Böttinger, "Speech is silver, silence is golden: What do asvspoof-trained models really learn?" *ArXiv*, vol. abs/2106.12914, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235624055

[13] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. S. Mello, R. V. Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi *et al.*, "Overview of btas 2016 speaker anti-spoofing competition," *In proceedings of IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, pp. 1–6, 2016.

[14] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv*, vol. arXiv:2109.00535, 2021.

[15] R. Baumann, K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik, "Voice spoofing detection corpus for single and multi-order audio replays," *Transaction on Computer Speech & Language*, vol. 65, p. 101132, 2021.

[16] I. Yakovlev, M. Melnikov, N. Bukhal, R. Makarov, A. Alenin, N. Torgashov, and A. Okhotnikov, "Lrpd: Large replay parallel dataset," *In proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6612–6616, 2022.

[17] A. N. Joon Son Chung and A. Zisserman, "Voxceleb2: Deep speaker recognition," *In proceedings of Interspeech*, pp. 1086–1090, 2018.

[18] https://github.com/TaoRuijie/ECAPA-TDNN.

[19] https://github.com/clovaai/voxceleb_trainer.

[20] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," *In proceedings of Interspeech*, pp. 2087—2091, 2015.

[21] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," *In proceedings of Interspeech*, pp. 1106—1110, 2020.

[22] 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 proceedings of the Speaker and Language Recognition Workshop (Odyssey)*, pp. 112–119, 2022.