Detecting Spoof Voices in Asian Non-Native Speech: An Indonesian and Thai Case Study

Aulia Adila, Candy Olivia Mawalim, and Masashi Unoki
Japan Advanced Institute of Science and Technology, Japan
E-mail: {adila, candylim, unoki}@jaist.ac.jp

***Abstract*—This study focuses on building effective spoofing countermeasures (CMs) for non-native speech, specifically targeting Indonesian and Thai speakers. We constructed a dataset comprising both native and non-native speech to facilitate our research. Three key features—MFCC, LFCC, and CQCC—were extracted from the speech data, and three classic machine learning-based classifiers—CatBoost, XGBoost, and GMM—were employed to develop robust spoofing detection systems using the native and combined (native and non-native) speech data. This resulted in two types of CMs: Native and Combined. The performance of these CMs was evaluated on both native and non-native speech datasets. Our findings reveal significant challenges faced by Native CM in handling non-native speech, highlighting the necessity for domain-specific solutions. The proposed method shows improved detection capabilities, demonstrating the importance of incorporating non-native speech data into the training process. This work lays the foundation for more effective spoofing detection systems in diverse linguistic contexts.**

I. Introduction

Over the past few years, voice-based authentication systems have become prominent for verifying identities and recognizing spoken utterances, offering convenience in various scenarios. However, they are vulnerable to spoofing attacks, particularly through logical access (LA) scenarios [1] involving Text-to-Speech (TTS) and Voice Conversion (VC). TTS generates natural-sounding artificial speech from text, while VC converts one speaker's voice to another's [2]. These techniques can create highly realistic speech, posing significant threats to the security and reliability of these systems.

The latest ASVSpoof 5 challenge [3] has demonstrated new surrogate detection models with adversarial attacks incorporated for the first time, using non-studio quality recordings that introduced a new challenge in the dataset. They built end-to-end systems using RawNet2 [4] and AASIST [5].

While much research has focused on spoofing detection in English and other languages [6], [7], there is a notable gap in studies addressing specific linguistic contexts, such as the diverse languages within the Asian region. English-speaking accents in Asian countries present unique challenges, as non-native accents make it difficult for spoofing detection systems to accurately differentiate between bonafide (genuine) and spoof speech.

This study focuses on Indonesia and Thailand due to Indonesia's large multicultural population and Thailand's unique dialect variability and tonal representation. Given the lack of publicly available datasets for these linguistic contexts, we developed an audio dataset with speech from Indonesian and Thai non-native English speakers. Our goal is to establish a spoofing countermeasure (CM) to effectively detect TTS and VC attacks for non-native speech accents and patterns.

The contribution of our work is summarized as follows:

- to facilitate the development of CMs to handle non-native speech by creating a dataset with non-native speech, addressing the gap in available speech datasets comprising non-native accents and characteristics,
- to understand the effectiveness on Native CMs in handling Asian non-native speech, and
- to propose CMs that can distinguish bonafide and spoof speech for both natives and non-natives.

In this study, we created a comprehensive dataset consisting of English native and Indonesian-Thai non-native speech to construct the CMs. Furthermore, we investigated how the CMs performed in distinguishing bonafide and spoof speech for both native and non-native speakers. Our findings show that the proposed method significantly improved CM performance in handling non-native speech. Additionally, our CMs utilize common front-end features and back-end classic machine learning (ML)-based classifiers, establishing a foundational baseline for this non-native speech study.

II. Related Works

In response to the growing threat of spoofing, the research community launched the ASVspoof challenges[1], a biennial initiative aimed at developing CMs to detect fake speech, starting in 2015 [8] and continuing through 2024 [3]. As these types of attacks have increased through the challenge editions, TTS and VC attacks have been included since the first edition in 2015.

Several CM systems have been developed to detect audio spoofing, primarily focusing on hand-crafted features that can effectively capture discriminative patterns of artifacts [9]. Early research mainly used conventional classifiers, such as ML-based classifiers.

A study on synthetic speech detection in 2015 found that dynamic LFCC with a GMM classifier performed best on ASVspoof evaluation sets [10]. In ASVspoof 2017 [11], the best system in the previous challenge was used as the baseline system. It was built using a common GMM back-end classifier with constant Q cepstral coefficients (CQCC). In ASVspoof 2019 [1], the top-performing systems employed LFCC features

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

with a light convolutional neural network (LCNN). However, other systems based on standard cepstral features and GMM-based classifiers were only slightly behind in performance. In ASVspoof 2021 [12], most systems operated using short-term spectral features or raw waveforms, utilizing ensemble systems and popular convolutional networks. These efforts primarily addressed spoofing in native language domains, leaving the non-native speech domain largely unexplored.

Speech corpora are essential for CM systems, and publicly available datasets with a large number of spoofing attacks have emerged to overcome data bottlenecks. These datasets, which include both bonafide and spoof speech generated by LA algorithms (TTS and/or VC), facilitate the evaluation and benchmarking of different systems. Notable datasets include ASVspoof2015 [8], ASVspoof2019-LA [13], ASVspoof2021-LA [12], and the latest ASVspoof2024 [3]. Other available datasets include WaveFake [14], ADD2022-LF [7], and Latin-American Voice Anti-spoofing [6]. However, datasets representing non-native English speech are scarce, highlighting the need for research focused on spoofing detection tailored to non-native accents.

III. Dataset Collection

The dataset used in this study consists of English native and non-native speech, spoken by Indonesian and Thai speakers. The dataset is split into training, validation, and testing sets with the ratio of 70-10-20 to ensure robust evaluation, with no overlap of speakers between the splits to maintain the integrity of the results. Table I summarizes our ENIT Dataset[2].

A. Bonafide Data

Our native English dataset consists of 7,990 utterances sourced from the training sets of ASVspoof 5 [3], which is derived from the English-language subset of the Multilingual Librispeech (MLS) dataset. The MLS dataset is a large multilingual corpus based on LibriVox audiobooks [15], featuring non-studio-quality recordings. This dataset includes recordings from over 4,000 speakers using various devices. As this dataset has a balanced number of speakers, we randomly selected 4,000 utterances from the total of 80 males and 80 females speakers and eliminated 10 utterances randomly to ensure the balance with our non-native English dataset.

The English non-native speech dataset consisted of 7,990 utterances recorded from 21 speakers, including 10 Indonesian speakers (7 males and 3 females) and 11 Thai speakers (7 males and 4 females). These speakers read articles sourced from online English newsletters covering unbiased topics such as health, astronomy, engineering, and technology. We selected similar news topics in Indonesian, English, and Thai, ensuring they did not express hate speech, violence, or harassment to maintain impartiality. Each news article was then segmented into sentences or sub-sentences with 5 to 20 words per utterance to ensure readability.

[2]ENIT Dataset: English Native and Indonesian-Thai Non-Native Speech

TABLE I
ENIT Dataset: English Native and Indonesian-Thai Non-Native Speech

Type	Sets	# unique spk.		# utterances		# spoof
		Male	Female	Bonafide	Spoof	attack
Native	Train.	56	56	5,590	33,540	8
	Dev.	8	8	800	4,800	8
	Eval.	16	16	1,600	9,600	8
Non-native	Train.	10 / 12	4 / 5	5,696	33,164	3
	Dev.	1 / 1	1 / 2	542	4,275	3
	Eval.	3 / 3	2 / 3	1,752	10,501	3

The recordings were conducted in a soundproof room using four different recording devices: a condenser microphone, a dynamic microphone, a mobile phone, and a laptop computer, without significant channel or background noise effects. All speech was recorded in a read style, with the speaking pace adjusted to each speaker, while still maintaining the naturalness of spontaneous speech. Audio post-processing included practical trim voice editing using Audacity[3] to separate the recordings by utterances.

B. Spoof Data

We constructed a corpus of 47,940 utterances by selecting a subset from the ASVspoof 5 training data [3]. The selected utterances are from the same 80 males and 80 females speakers used in the bonafide data. The number of spoof utterances is six times larger than the number of bonafide utterances. Table I specifies the number of unique speakers of bonafide and spoof speech on the left and right side of the separator '/', respectively. If the number of speakers is the same for both bonafide and spoof speech, we do not use the separator.

To generate spoof data for non-native English speech, we employed TTS, VC, and synthesis techniques. These methods can be broadly categorized into three distinct approaches.

1) SpeechT5 [16]: SpeechT5 is a model that unifies speech and text processing under a single framework. Inspired by the success of T5 in the text domain [17], SpeechT5 employs an encoder-decoder architecture to learn shared representations for both modalities. It incorporates specialized pre- and post-processing networks to handle the unique characteristics of speech and text, respectively.

Through extensive self-supervised pre-training on large-scale unlabeled speech and text data, SpeechT5 acquires a deep understanding of both modalities. This versatility empowers it to excel in a wide range of tasks, including automatic speech recognition, speech synthesis, speech translation, VC, speech enhancement, and speaker identification.

Approximately 5,200 synthetic utterances were generated using the SpeechT5 model and x-vector speaker embeddings [18]. To manipulate acoustic and linguistic characteristics, we varied both x-vector and bottleneck features (representing linguistic information). Roughly 2,000 utterances were created

[3]Audacity (ver. 3.4.2) is an open source software for recording and editing sounds.

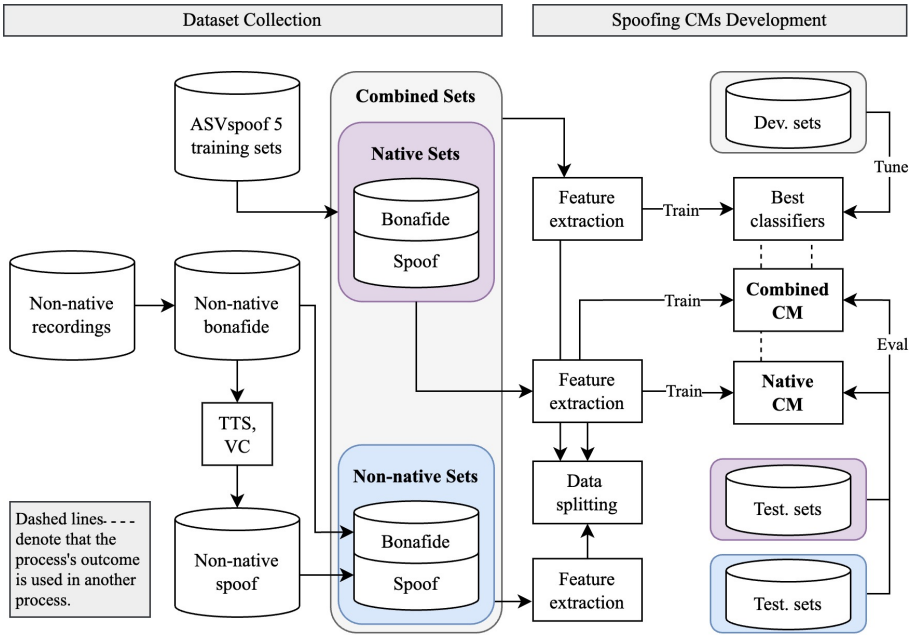

Fig. 1. Block diagrams of our proposed method for building spoofing CMs

using x-vector and bottleneck features extracted from our collected bonafide non-native speaker data. In another set of 1,000 utterances, x-vectors from native speakers in the CMU ARCTIC dataset [19] were combined with bottleneck features from our non-native speakers. The remaining utterances were synthesized by pairing x-vectors from our non-native speakers with bottleneck features from the VCTK dataset [20].

2) FreeVC [21]: We utilized FreeVC, a text-free VC system capable of transforming a speaker's voice with only a single reference audio sample. The model builds upon the ViTS architecture [22] but departs from traditional text-based approaches by learning to disentangle content information directly from raw audio waveforms. FreeVC extracts speaker-independent features that are subsequently processed through a bottleneck layer to isolate content representations using WavLM [23]. To further enhance the model's ability to separate content from speaker identity, spectrogram resizing is employed as a data augmentation technique.

Central to the one-shot conversion capability is a speaker encoder that extracts speaker-specific characteristics. We experiment with two configurations of the speaker encoder: FreeVC-s, which utilizes a non-pretrained encoder, and FreeVC, which benefits from a pre-trained encoder. This comparative analysis allows us to assess the impact of pre-training on the overall performance of the VC system.

3) WORLD [24]: We also employed the WORLD, a vocoder-based speech synthesis system, to generate synthetic spoof data. This system is widely recognized for its exceptional performance in speech analysis, manipulation, and synthesis. It analyzes speech into three parts: fundamental frequency (F0), aperiodicity, and spectral envelope.

To further refine the spectral envelope estimation process, we integrated the CheapTrick algorithm [25] into our pipeline. This method leverages F0-adaptive windowing to enhance spectral resolution. By smoothing the power spectrum and applying spectral recovery techniques in the quefrency domain, CheapTrick better estimates the spectral envelope. This improved accuracy contributes significantly to the overall quality of the synthesized speech. We manipulated speed and F0 with random parameters during the speech synthesis process. This randomization helps to create a more diverse and challenging dataset for subsequent anti-spoofing models.

IV. Proposed Method

A block diagram of our proposed method in building the spoofing CMs using our datasets is shown in Figure 1. To establish a foundation for our newly developed dataset, which represents the characteristics of non-native speech, we employed hand-crafted features and classic ML-based classifiers.

A. Feature extraction

The front-end features are mostly derived from the magnitude or power spectrum, encompassing both short-term and long-term magnitude spectral features [9], to effectively capture the essential characteristics of speech signals, including formant structures and energy distribution across different frequency bands. These features are typically obtained from MFCC [26], LFCC [27], and CQCC [28].

LFCC is widely used in speaker recognition and has demonstrated strong performance in spoofing detection. MFCC has been extensively explored for accent classification tasks, which closely relate to this study on non-native speech. CQCC has also proven effective in detecting audio spoofing by providing higher frequency resolution at lower frequencies and higher temporal resolution at higher frequencies. This study aims to investigate the effectiveness of LFCC, MFCC, and CQCC

in detecting spoofing attacks in non-native English speech, focusing on Indonesian and Thai accented speakers.

B. Classifiers

We use classic ML-based classifiers to build the baseline spoofing CMs. To determine the most effective classifier, we conducted model selection on several widely used ML-based classifiers and evaluated their performance on our development sets. On the basis of our experiments, we identified the two best classifiers for the features used (MFCC, LFCC, CQCC): CatBoost [29] and XGBoost [30].

CatBoost is a type of binary decision trees-based gradient-boosting predictor that has proven effective in dealing with categorical data [31]. Categorical data consists of distinct values that cannot be compared directly. In this study, we deal with binary categorical data.

Extreme Gradient Boosting, or XGBoost, is a gradient-boosted tree algorithm for supervised learning. Gradient-boosted decision trees excel in learning from noisy data and have achieved cutting-edge results. XGBoost is a gradient-boosting implementation that optimizes computing performance to provide a scalable solution.

Additionally, we included the GMM-based classifier as our baseline, which has been a fundamental method in the ASVspoof challenge series. GMM assumes that data points belong to a mixture of a finite Gaussian distribution [32]. It is commonly used in anti-spoofing systems. There, separate GMMs are learned for each bonafide and spoof dataset. The classification of new input is predicted by calculating the ratio of its log-likelihood of belonging to bonafide and spoof GMMs.

V. Experimental Settings

As the currently available speech corpora primarily consist of native speech, we aim to assess the performance of CMs built upon these datasets in handling non-native speech. We refer to this CM as Native CM. Subsequently, we developed a spoofing CM specifically for non-native speech by combining the current publicly available datasets, which are predominantly native speech, with our non-native datasets. We refer to this CM as Combined CM. Our experiment pipeline is illustrated in Fig. 1.

A. Experiment 1: Assessing the Native CM on Non-Native Speech

First, we built the Native CM by training our selected models (GMM, CatBoost, and XGBoost) using the extracted features (MFCC, LFCC, and CQCC) on the native speech training sets. The features were extracted using the Matlab toolbox and Smileslab implementations [33]. We set the random seed to 42 for all classifiers and adjusted the hyperparameters as detailed in Subsection V-C. Each model was then evaluated on both native and non-native speech evaluation sets, using the specified evaluation metrics. This initial assessment on native speech helps us understand the model's performance in a typical native speech context.

TABLE II
CMs performance evaluation result in Experiments 1 and 2

Exp.	Feat.	Classifier	Non-native		Native	
			minDCF	EER (%)	minDCF	EER (%)
1	MFCC	CatBoost	0.81	41.54	**0.17**	**6.88**
		XGBoost	0.83	42.58	0.18	7.48
		GMM	1.00	100	1.00	100
	LFCC	CatBoost	0.95	40.07	0.24	10.19
		XGBoost	0.98	40.36	0.26	10.56
		GMM	1.00	62.84	0.83	33.89
	CQCC	CatBoost	0.84	38.11	0.29	12.43
		XGBoost	0.79	35.57	0.28	12.56
		GMM	1.00	76.02	0.97	47.69
2	MFCC	CatBoost	0.33	12.66	0.19	7.81
		XGBoost	0.38	13.98	0.21	8.08
		GMM	0.96	46.58	1.00	46.18
	LFCC	CatBoost	0.27	10.90	0.31	13.32
		XGBoost	0.25	10.34	0.33	13.76
		GMM	0.98	53.31	1.00	46.13
	CQCC	CatBoost	**0.19**	**8.56**	0.37	15.06
		XGBoost	0.21	9.63	0.36	14.75
		GMM	0.97	41.55	0.99	91.89

B. Experiment 2: Assessing the Combined CM on Non-Native Speech

With regard to the Combined CM, we trained our selected models (GMM, CatBoost, and XGBoost) using the extracted features (MFCC, LFCC, and CQCC) on the combined native and non-native speech training sets. We set the same hyperparameter tuning as experiment 1 to ensure fair comparison. We also used the same evaluation protocol to facilitate the common assessment and benchmarking of both CMs.

C. Hyperparameter settings

We set different hyperparameters for each classifier on the basis of its architecture as implemented in the scikit-learn library [32]. For the GMM classifier, we used two mixture components, a full covariance type, the k-means method for initializing the weights, and a maximum of 100 iterations. For CatBoost, we set the verbose parameter to 0. For XGBoost, we used logarithmic loss as the evaluation metric. All other hyperparameters were kept at their default configurations.

D. Evaluation Metrics

In this study, we used minimum detection cost function (minDCF) and equal error rate (EER) adopted from the latest ASVspoof 5 challenge, specifically on track 1: stand-alone spoofing and speech deepfake detection [3]. The spoofing detection employed in the challenge is built upon the normalized detection cost function (DCF), defined as follows:

$$\mathrm{DCF}'(\tau_{\mathrm{cm}}) = \beta \cdot P_{\mathrm{cm,\ miss}}(\tau_{\mathrm{cm}}) + P_{\mathrm{cm,\ fa}}(\tau_{\mathrm{cm}}) \tag{1}$$

where

$$\beta = \frac{C_{\mathrm{miss}} \cdot (1 - \pi_{\mathrm{spf}})}{C_{\mathrm{fa}} \cdot \pi_{\mathrm{spf}}} \tag{2}$$

In Eq. (1), $P_{\mathrm{cm,\ miss}}(\tau_{\mathrm{cm}})$ and $P_{\mathrm{cm,\ fa}}(\tau_{\mathrm{cm}})$ are the empirical miss rate for bonafide utterances and false alarms for spoof utterances, respectively. Both are regarded as a function of the detection threshold τ_{cm}. The constant β In Eq. (2) is calculated

using miss rate cost C_{miss}, false alarm cost C_{fa}, and detection threshold τ_{cm} with the values set as follows: $C_{\text{miss}} = 1$, $C_{\text{fa}} = 10$, and $\pi_{\text{spf}} = 0.05$. Therefore, the value of β is approximately 1.90.

Furthermore, Eqs. (1) and (2) are used to compute the minimum DCF (minDCF), which measures the CMs performance by using the threshold set on the basis of ground-truth where the lower value indicates a better performance. This metric is calculated as follows:

$$\text{minDCF} = \min_{\tau_{\text{cm}}} \text{DCF}'(\tau_{\text{cm}}) \tag{3}$$

The final evaluation metric used is the EER, one of most widely used metrics for audio spoofing CMs. It represents the CM threshold where the false acceptance rate equals the false rejection rate. A lower EER value signifies better performance.

VI. Results

As seen in Table II, the GMM classifier has comparably high minDCF and EER values on the native speech evaluation sets. It also reached the maximum possible DCF and EER values of 100% on MFCC features, indicating extremely poor performance as it is entirely ineffective at distinguishing between bonafide and spoof speech in both non-native and native contexts. The lowest EER for the GMM classifier was achieved with LFCC features, which is expected as this system is considered the baseline in the ASVspoof challenge.

While the CatBoost and XGBoost classifiers achieved competitive results on native datasets, they still obtained minDCF values greater than 0.79 and EER values higher than 38% on non-native datasets. All the CMs in experiment 1 on average declined around 0.44 in minDCF and 19% in EER relative to their performance on native datasets. These results indicate a significant lack of non-native speech handling capability in Native CMs, demonstrating the inefficacy of the pre-existing speech spoofing CMs in distinguishing between bonafide and spoof speech in non-native contexts.

In experiment 2, all Combined CMs significantly improved when predicting the non-native datasets, with an average relative improvement of around 0.41 in minDCF and around 30% in EER. However, this experimental setting slightly worsened the CMs' ability to detect native speech, as indicated by an average of total increase on minDCF and EER around 0.06 and 1.7%, respectively. This behavior may be due to the increased variability introduced by the non-native speech data within the training sets, making it more challenging for the model to generalize well to the specific characteristics of native speech alone, as it now has to account for a broader range of speech patterns.

Among all the CMs we built, CQCC feature extraction with CatBoost classifier performed the best when evaluated on non-native datasets. These ensemble models tended to perform better due to their boosting algorithm, which iteratively builds an ensemble by training each new model to correct the errors of the previous ones.

Meanwhile, the GMM classifier, which is often used as a baseline model, performed worse than other classifiers. In most cases, the GMM classifier predicted the input speech as a spoof. Despite reducing EER by up to 50% after being trained using the combined native and non-native datasets, its performance as a CM is not satisfactory.

VII. Conclusions

In this study, we constructed spoofing countermeasures (CMs) to handle non-native speech using different acoustic features (MFCC, LFCC, and CQCC) and classic machine learning-based classifiers (CatBoost, XGBoost, and GMM). We also developed the ENIT dataset, which includes speech from Indonesian and Thai non-native English speakers. We conducted two experiments to assess the current CMs (Native CM) in handling non-native speech, as well as the baseline we constructed (Combined CM).

Our findings show that the Combined CMs is better at detecting bonafide and spoof non-native speech than the Native CMs, improving minDCF and EER scores around 0.41 and 30% on average, respectively. Moreover, we also found that CatBoost and XGBoost obtained competitive results using all three features, while GMM performed the worst in all experiments. In our future work, we will expand the datasets to be able to capture more non-native speech from the Asian region, as well as building more accurate and robust CMs focusing on accent and other speech characteristics.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), Grant-in-Aid for Transformative Research Areas (A) (23H04344). This work was a part of the ASEAN IVO project titled 'Spoof Detection for Automatic Speaker Verification'(www.nict.go.jp/en/asean_ivo).

References

[1] A. Nautsch, X. Wang, N. Evans, *et al.*, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," *IEEE Trans. on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, Apr. 2021, ISSN: 2637-6407.

[2] M. Sahidullah, H. Delgado, M. Todisco, *et al.*, "Introduction to Voice Presentation Attack Detection and Recent Advances," *CoRR*, vol. abs/1901.01085, 2019.

[3] X. W. et al., "ASVspoof 5: Crowdsourced data, deepfakes and adversarial attacks at scale," in *ASVspoof 2024 workshop (submitted)*, 2024.

[4] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. W. D. Evans, and A. Larcher, "End-to-End anti-spoofing with RawNet2," in *Proc. of ICASSP 2021*, IEEE, 2021, pp. 6369–6373.

[5] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," in *Proc. of ICASSP 2022*, 2022, pp. 6367–6371.

[6] P. A. Tamayo Flórez, R. Manrique, and B. Pereira Nunes, "HABLA: A Dataset of Latin American Spanish Accents for Voice Anti-spoofing," in *Proc. of INTERSPEECH 2023*, 2023, pp. 1963–1967.

[7] J. Yi, R. Fu, J. Tao, *et al.*, "ADD 2022: the first Audio Deep Synthesis Detection Challenge," in *Proc. of ICASSP 2022*, IEEE, 2022, pp. 9216–9220.

[8] Z. Wu, T. Kinnunen, N. Evans, *et al.*, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. of INTERSPEECH 2015*, 2015, pp. 2037–2041.

[9] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio Anti-Spoofing Detection: A Survey," *CoRR*, 2024. arXiv: 2404.13914 [cs.SD].

[10] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. of INTERSPEECH 2015*, ISCA, 2015, pp. 2087–2091.

[11] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. of INTERSPEECH 2017*, 2017, pp. 2–6.

[12] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," *CoRR*, vol. abs/2109.00537, 2021.

[13] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, p. 101 114, 2020.

[14] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," *CoRR*, vol. abs/2111.02813, 2021.

[15] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech 2020*, ser. interspeech$_2$020, ISCA, Oct. 2020.

[16] J. Ao, R. Wang, L. Zhou, *et al.*, "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing," in *Proc. of ACL 2022*, Association for Computational Linguistics, 2022, pp. 5723–5738.

[17] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, 140:1–140:67, 2020.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. of ICASSP 2018*, IEEE, 2018, pp. 5329–5333.

[19] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA ITRW on Speech Synthesis, Pittsburgh, PA, USA, June 14-16, 2004*, A. W. Black and K. A. Lenzo, Eds., ISCA, 2004, pp. 223–224.

[20] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)*, 2019.

[21] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion," in *Proc. of ICASSP 2023*, IEEE, 2023, pp. 1–5.

[22] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. of ICML 2021*, vol. 139, PMLR, 2021, pp. 5530–5540.

[23] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99-D, no. 7, pp. 1877–1884, 2016.

[25] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE TASLP*, vol. 28, no. 4, pp. 357–366, 1980.

[27] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[28] H. Tak, J. Patino, A. Nautsch, N. W. D. Evans, and M. Todisco, "An Explainability Study of the Constant Q Cepstral Coefficient Spoofing Countermeasure for Automatic Speaker Verification," in *Proc of Odyssey 2020*, K. Lee, T. Koshinaka, and K. Shinoda, Eds., ISCA, 2020, pp. 333–340.

[29] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Proc. of NeurIPS 2018*, 2018, pp. 6639–6649.

[30] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. of ACM SIGKDD 2016*, ser. KDD '16, ACM, Aug. 2016.

[31] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *CoRR*, vol. abs/1810.11363, 2018.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward," *CoRR*, vol. abs/2210.00417, 2022.