APSIPA Transactions on Signal and Information Processing, 2025, 14, e14

Original Paper

A Brain-inspired Multi-Detector Machine for Fake Speech Detection

Chang Feng[1], Xiaolong Wu[1,2], Mingxing Xu[1] and Thomas Fang Zheng[1*]

[1] *Center for Speech and Language Technologies, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China*

[2] *School of Computer Science and Technology, Xinjiang University, Urumqi, China*

ABSTRACT

Fake speech detection is essential to defend speech spoofing attacks of imitation through Artificial Intelligence Generated Content (AIGC) technologies including text-to-speech synthesis (TTS) and voice conversion (VC). Existing solutions are classification-based and rely on large data, showing limitations in solving both data diversity and result explainabilty problems. To resolve these limitations, we design a Brain-inspired Multi-Detector Machine (BiMDM), which is inspired by the brain's perception and decision-making mechanisms. Our method proposes to use multiple detectors to capture various aspects of fake speech characteristics. To ensure the final detection precision, each detector is trained with the aim of Maximum Detection Precision (MDP) for a specific forgery clue, unlike previous classifiers optimized for Minimum Classification Error (MCE). And a sufficient number of detectors are necessary to reduce the total detection miss rate. This mechanism assigns meaningful roles to each individual detector as well as ensures the detection performance. Then the detectors' results are integrated through an overall explainable decision-making module, including OR logic calculus and decision trees, to produce result

*Corresponding author: Thomas Fang Zheng, fzheng@tsinghua.edu.cn.

Received 05 March 2025; revised 25 April 2025; accepted 22 May 2025
ISSN 2048-7703; DOI 10.1561/116.20250011

with explainability of the entire detection process. Our experimental results demonstrate the effectiveness of our multi-detector machine and reveal the potential of our proposed novel perspective for fake speech detection task.

Keywords: Fake speech detection, multiple detectors, decision trees, OR logic

1 Introduction

Fake speech detection is to defend fake speech spoofing attacks from Artificial Intelligence Generated Content (AIGC) technologies including text-to-speech synthesis (TTS) [3, 16] and voice conversion (VC) [9, 19]. As these technologies continue to improve, fake speech has become increasingly accessible and closer to genuine human speech, posing significant security risks for voice authentication systems, audio recording evidence authentication and telephone communication. Therefore, the importance of fake speech detection is growing. Currently, the types of fake speech generation algorithms are diverse, which requires the fake speech detection methods to handle data with diversity. Additionally, the reliability of AI-based detection is also crucial in practical security applications, which requires fake speech detection methods to be able to provide explanations for their results.

Existing methods treat the fake speech detection problem as a classification task, typically a binary classification problem to distinguish between fake and genuine speech. These approaches employ classifiers [8, 10, 30] optimized for Minimum Classification Error (MCE), which calculate a confidence score indicating the likelihood of the input data belonging to a specific class. Such classification-based methods need to consider the characteristics of both classes and rely on large data for training, and when there is data diversity within one class (e.g., the fake speech class), it becomes more difficult to capture the distinctions between the two classes as well as the commonalities within the same class. Additionally, the results of classifiers are derived from balancing the two classes, and this balancing process is opaque and fails to provide explanations aligning with human reasoning and decision-making logic. Therefore, the existing methods still show limitations in solving both data diversity and result explainabilty problems.

To resolve these limitations and problems, this paper proposes a novel approach that constructs global detection information from multiple individual detection clues, moving away from the traditional classification perspective. We design a Brain-inspired Multi-Detector Machine (BiMDM), which is inspired by the brains perception and decision-making mechanisms. In the BiMDM, each detector is independent of each other. And each is assigned to

detect whether a specific forgery clue is shown in the speech audio, with 100% confidence, which is designed with the goal of Maximum Detection Precision (MDP). It is implemented by setting threshold for the computation results and adding precison penalty in the training process. The output of a detector is a binarized value of bool type, where true value indicates that the forgery clue has been detected, and false value indicates that it has not been detected. And a sufficient number of detectors are necessary to reduce the total detection miss rate, which ensures the detection performance. The detectors' results are then integrated through an overall explainable decision-making module with decision trees concatenated with OR logic calculus, which maintains fault tolerance to detection errors of detectors. Such decision-making module is similar to human reasoning in decision-making process and can provide explanations for the final detection result.

This paper evaluates the performance on ASVspoof2019 Logical Access (LA) dataset and ASVspoof2021 LA dataset. We analyze the overall detection performance as well as the breakdown performance according to different fake speech generation algorithms and signal distortion conditions, comparing with the two state-of-the-art methods mentioned on the dataset's official website. Furthermore, we also trace back to the detectors' results, with which the final detection result obtains explanations. The experiments demonstrate that our method is capable of handling various fake speech types and providing explanations, which reveals the potential of our proposed novel perspective for fake speech detection task.

The rest of the paper is organized as follows. We reviewed related work in Section 2. BiMDM design details are described in Section 3. Our experimental details and the evaluation results are reported in Section 4. The conclusion of our work is in Section 5.

2 Related Work

Previous methods have primarily focused on developing feature extraction techniques as well as classifier models to improve the detection performance. The features are extracted from spectral-based information [11, 24], phase-based information [25, 37] or learned through deep learning methods [20, 36, 38, 39]. As for classifier model design, there are traditional machine learning models [26, 27], Convolutional Neural Network (CNN) models [36], Deep Residual Network (ResNet) models [15], Graph Attention Network (GAT) models [28] and end-to-end models [7, 30, 33] that directly operate on audio sampling points.

To further handle various fake speech data in fake speech detection task, researchers explore to fuse multiple aspects of information. The multiple aspects of information can be obtained by extracting different features or em-

ploying different models to capture various aspects of the data. Pal *et al.* [21] proposes to use three kinds of features, including prosody information, spectral-based features and phase-based features to construct three different classifiers, and then perform a weighted sum of the scores from the three classifiers. Kumar *et al.* [14] propose to input prosody information into four types of classifiers and then obtain the final results by averaging the output scores from the four classifiers. Fan *et al.* [4] propose to use different information of complex spectral features, including Log Power Spectrogram (LPS), real and imaginary spectrograms to build different DNN-based classifiers and using a weighted sum of scores to obtain the final resutls. Tak *et al.* [29] extracts different sub-band information for sub-band classifiers and then integrates them through non-linear computation. There are also approaches of multi-path models [5, 35] to learn several models on different speech data and then integrate the models. But in one word, the existing information fusion approaches are based on classifier models from the classification perspective, which focuses on fusion of score from the classifiers.

3 Brain-inspired Multi-Detector Machine Design

The process through which the human brain makes decisions based on external information is a complex, layered mechanism that begins with perception. This perception relies on multiple specialized sensory receptors [22], each specialized to detect a specific type of environmental stimuli. These receptors process information to a binarized form with the state of activation or inactivation. If the stimuli match the detection pattern, specific neural pathways are activated to further process and respond to the information. In contrast, when the stimuli do not meet the required criteria, the neural pathways remain inactive. Additionally, each receptor is assigned specific meaning and function within the brain's processing system, which allows for distinct recognition of environmental cues. The brain then integrates the sensory inputs from various receptors into a unified global representation of the environment. This integrated perception serves as the foundation for further logical reasoning and ultimately guides decision-making processes, ensuring that actions are based on a comprehensive understanding of the surroundings.

Inspired by the multi-clue perception and decision-making mechanism, we propose BiMDM for fake speech detection task where the fake speech audios are diverse and possess many aspects of forgery clues [6, 13, 27]. Our proposed BiMDM is structured in two key stages: the artifact detection stage and decision making stage, as shown in Figure 1. The two stages are trained independently. The first stage processes the speech audio input through multiple detectors, each specialized to identify different forgery clues related to four specific artifact aspects. These detectors function similarly to sensory re-

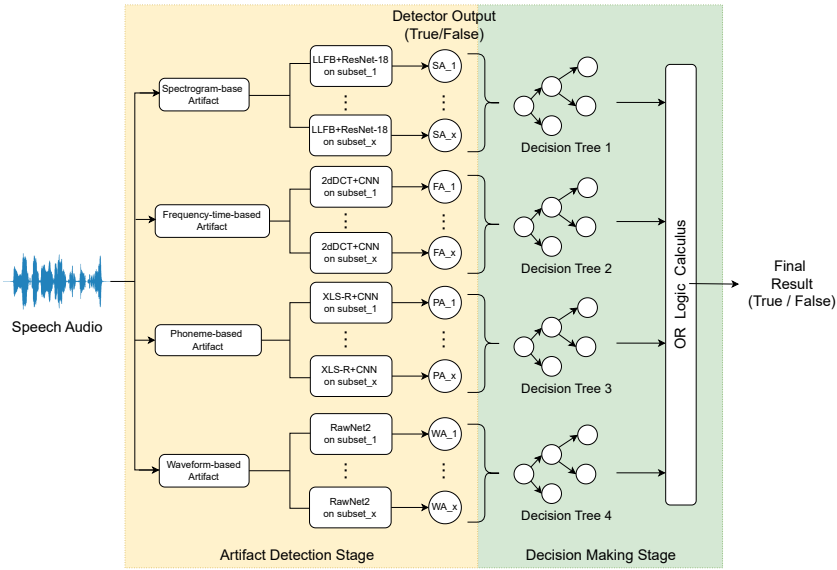


Figure 1: The overview of our proposed Brain-inspired Multi-Detector Machine (BiMDM).

ceptors in the human brain, where each detector outputs a binarized value for its state of activation or inactivation and the forgery clue pattern is matched with maximized precision. The aim of detectors is different from the MCE aim of classifiers in previous methods. The second stage solely relies on the outputs from the first stage, combining these detection results to generate the final task result. Unlike previous model-fusion methods that focus on score-level fusion with weighted sum of scores from the classifiers' results, our BiMDM works with binarized outputs of detectors. And its decision-making structure with decision trees mimics the logical inference of the human brain.

This method can be formulated in the form of mathematical set, as in Figure 2. We define S to be the universal set containing all the targets (the entire dataset of fake speech). There are n detectors with D_1, D_2, ..., D_n. In theory, each detector D_i detects a subset $S_i \subseteq S$ perfectly with 100% precision and no non-targets (i.e. genuine speech) will be detected. So the non-targets will never be detected and can be disregarded in the detection set. Our task goal is simplified to covering the entire set S using the union of the sets detected by all n detectors. In other words, every element in S should be detected by at least one of the detectors. The coverage formulation can be presented by $S \subseteq \bigcup_{i=1}^{n} S_i$, where S_1, S_2, ..., S_n are the subsets of S that each detector D_1, D_2, ..., D_n detects respectively. To cover the entire set S,

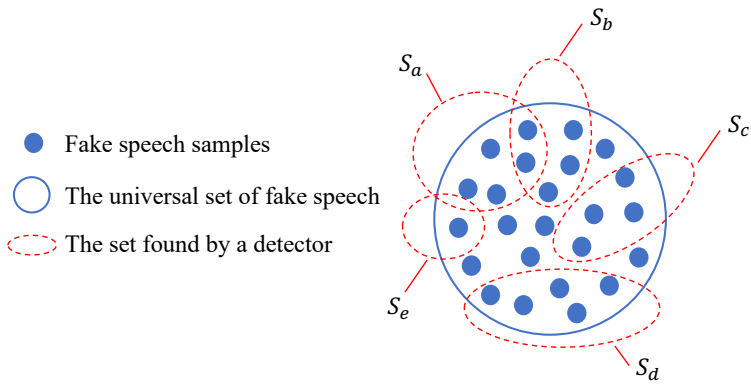


Figure 2: Fake speech detection problem in the form of mathematical set. In BiMDM, the detectors with maximum precision check out only fake speech data, so we consider only fake speech data set. Here, the relationship of subset S_a and S_b is $S_a \cap S_b \neq \emptyset$, while the relationship of subset S_a and S_c is $S_a \cap S_c = \emptyset$.

we require that every element of S is in at least one of these subsets, which is guaranteed if S is a subset of the union of all S_i. Additionally, the subset of S is independent of each other. These subsets may either intersect or be disjoint, meaning that there is no restriction on whether their pairwise intersections are non-empty or empty. The independence refers to the lack of any dependency or specific relationship between the subset S_i and S_j, where $i \neq j$, regardless of whether they share common elements. That is, $S_i \cap S_j = \emptyset$ or $S_i \cap S_j \neq \emptyset (i \neq j)$ both meet the conditions of detection. In the meanwhile, for subset S_i detected by the detector D_i, its uncovering subset is not considered for the detector D_i.

3.1 Artifact Detection Stage

This stage composes of multiple detectors that focus on a specific forgery clue and handles a portion of fake speech detection information with detection precision of 100% in probability.

A detector comprises a computing model that calculates information patterns and a threshold θ_{opt} that can result in the highest detection precision. By comparing the score of the detection computing model with threshold θ_{opt}, the output of a detector is obtained and is in the form of binarized bool value, indicating whether the forgery clue is present or not. If the score exceeds the threshold, the response is true; otherwise, it is false. The detector output $D(x)$ on sample x with computing score $\hat{y}$ can be formulated as

$$D(x) = \begin{cases} \text{True}, & \text{if } \hat{y} > \theta_{opt} \\ \text{False}, & \text{if } \hat{y} \leq \theta_{opt} \end{cases} \tag{1}$$

Threshold Search

The threshold θ_{opt} is determined through a process called *ThresholdSearch*, which can be formalized as follows:

$$\theta_{opt} = \arg\max_{\theta} Precision(\theta) \quad subject\ to \quad \theta \in [\theta_{min}, \theta_{max}]. \tag{2}$$

where

- $Precision(\theta)$ is the detection precision of the detector when the threshold Th is applied on the current dataset (i.e. training or development dataset).
- θ_{min} and θ_{max} represent the minimum and maximum possible threshold values.

Training for Maximum Precision

The detectors in the BiMDM are designed with the aim of MDP, that is, the detection precision of 100% in probability. It needs to pay more attention to attributes of the target samples (i.e. fake speech samples) in the training process. And as the output of detector is a binarized value that relies on the fixed decision threshold obtained through $Threshold\ Search$ on the training or development dataset, we design a detection precision penalty term in addition to the traditional binary cross-entropy loss ($\mathcal{L}_{BCE}$) which is typically for MCE aim for two classes of samples and without the need of decision threshold. The detection precision penalty term is defined as

$$P_{penalty} = y \cdot ReLU(\theta^{Train}_{p_{train}} - \hat{y}), \tag{3}$$

where

- y is the label of the sample (the fake speech set to 1 and the genuine to 0) and $\hat{y}$ is the model's computing score.
- $\theta^{Train}_{p_{train}}$ is the threshold obtaining Precision of p_{train} calculated on the training dataset through *ThresholdSearch*. Here p_{train} is set to very close to 100% for approaching the maximum precision aim.
- ReLU is Rectified Linear Unit which is an activation function in neural networks that outputs the input directly if it is positive, and zero otherwise.

Such penalty term makes an extra penalization on the false prediction of target samples (those with fake speech label of 1) under the condition of searched threshold $\theta^{Train}_{p_{train}}$. If the computing score for fake speech samples is lower

than the threshold, the penalty increases as the score deviates further below the threshold. This term enhances attention to the target samples. And by considering the decision threshold for the binarized output in the training, the detection with the fixed decision threshold becomes more robust.

Then, the training loss function is

$$\mathcal{L} = \mathcal{L}_{BCE} + \lambda P_{penalty}, \tag{4}$$

where λ is the penalty term parameter that controls the strength of the detection precision penalty and is set to a value greater than zero to produce the desired effect.

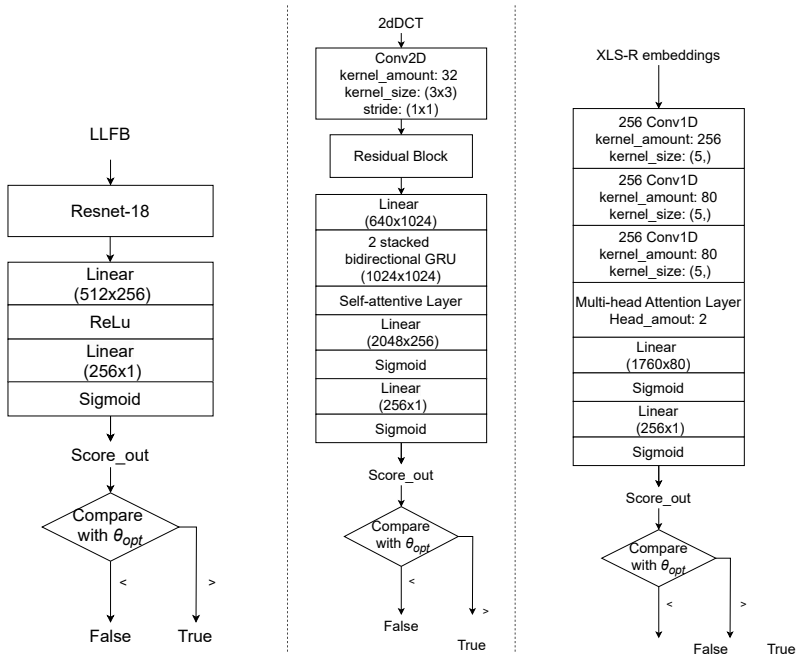
Detectors for Fake Speech Artifacts

In BiMDM, the detectors are independently with each other. Each detector can be configured with different inputs or computational structures, depending on the specific forgery clue that it is designed to capture. This flexibility allows the detectors to be tailored for detecting various types of clues. Since different fake speech generation algorithms introduce different forgery clues, we follow the multi-path strategy [35, 5] to learn several individual detectors respectively on fake speech of different generation algorithms.

Here, we consider artifacts from four kinds of domain, including spectrogram, frequency-time, phoneme, and time aspects, and applying different computing model structures to capture them.

- **Detector for spectrogram domain artifact (SA).** Forgery clues of fake speech can be observed in the spectrogram, including overly smooth spectral patterns, where high-frequency details are often lacking, making the speech sound muffled or unnatural. There may also be spectral gaps or attenuation in different sub-band frequency ranges, affecting clarity and sharpness. Additionally, the frequency bands may appear overly focused or concentrated in specific ranges, especially during phoneme transitions, unlike the broader distribution seen in genuine speech. We extract log-linear filter bank (LLFB) features and a pre-trained ResNet-18 [11] to represent the spectrogram information. The computing model is two linear layers with a ReLU activation layer. This structure detail is shown in Figure 3a.

- **Detector for frequency-time domain artifact (FA).** Forgery clues of fake speech related to both the frequency and time domains include mismatches or misalignments between the spectral and time components. In genuine speech, the frequency components dynamically change in sync with physiological pronunciation process. But in fake speech, certain frequency components may not synchrize with the time, appearing



(a) (b) (c)

Figure 3: Detail structures of computation models for (a) spectrogram domain artifact, (b) frequency-time domain artifact, (c) phoneme domain artifact. Here, the value of each detector's θ_{opt} is different.

at the wrong time or failing to appear when expected. We apply 2-Dimensional Discrete Cosine Transform (2dDCT) which first performs DCT in the time domain to extract the temporal characteristics of the speech signal, and then another DCT in the frequency domain to compress the frequency redundancy within each column and extract features related to the formant group. The computing model is CNN-based [6] with detail in Figure 3b.

- **Detector for phoneme domain artifact (PA).** As forgery clues can be found in phoneme pronunciation, which are the smallest meaningful units of sound in speech. The clues include lack of subtle variation that occurs in natural speech due to the complex coordination of the

tongue, teeth, throat, airflow, and vocal cords. Fake speech often exhibits a mechanical quality, especially in phonemes that require precise oral adjustments or changes in airflow. This can lead to distorted or blurry pronunciation, especially for certain consonants or vowels, which may sound unnatural or imprecise. So we utilize XLS-R (Cross-Lingual Speech Representation) [1], a powerful pre-trained model designed for robust speech recognition across multiple languages to generate representation of phoneme pronunciation. The computation model is CNN-based with Multi-head Attention mechanism, the detail structure is shown in Figure 3c.

- **Detector for time domain artifact (WA).** Forgery clues in fake speech can be detected through irregularities in the distribution and smoothness of the signal sampling points across time. Unlike genuine speech, which maintains continuous with densely packed and smooth sampling points reflecting natural vocal fluctuations, fake speech often exhibits noticeable discretization and abrupt changes due to limitations in generation algorithms, especially those based on neural network. We utilize RawNet2 [30] model with SincNet [23] with 1024-sample filter length as the computing model operating directly on raw audio data.

3.2 Decision Making Stage

To integrate the outputs from each detector, we employ a sequential structure that mimics the process of human reasoning.

One strategy is to apply OR logic calculus to the detectors' outputs of binarized bool value, which represents the simplest simulation of the inference process. In this method, if any forgery clue is detected, the inference decision is made that the speech is fake.

Another strategy is to first categorize the detectors' outputs according to the types of artifact. That is, the detectors capturing the same artifact detection trained with multi-path strategy belong to the same category. Within the artifact category, the detectors' outputs are organized in a decision tree, which ensures that the decision-making process aligns with logical reasoning and enhances explainability. As we employ four types of artifacts, there are four decision trees in the decision-making process. The decision trees are finally connected by OR logic calculus.

4 Experiments

4.1 Dataset

The experiments were conducted on two publicly available datasets: the ASVspoof2019 Logical Access (LA) dataset [34] and the ASVspoof2021 LA dataset [18], which are both specifically designed for the speaker verification spoofing caused by fake speech from TTS and VC techniques.

The ASVspoof 2019 LA dataset (19LA) includes clean speech data with genuine speech from Voice Cloning Toolkit (VCTK) corpus [32] and nineteen types of fake speech from TTS and VC techniques. It is divided into three mutually exclusive subsets: the training set, which is used for model training; the development set, for hyperparameter tuning; and the evaluation set, which serves for model validation. The spoofed speech in the training and development sets is generated by six different generation algorithms (labeled A01 to A06), while the evaluation set contains spoofed speech from the other thirteen algorithms (labeled A07 to A19).

In contrast, the ASVspoof 2021 LA dataset (21LA) introduces additional complexity by simulating more realistic conditions, such as signal encoding and transmission distortions typically encountered in real-world communication environments. While the 19LA dataset contains clean speech data under controlled conditions (without noise, reverberation, or channel distortions), the 21LA dataset considers the effects of communication transmission and encoding. Specifically, the 21LA dataset builds upon the 19LA dataset by applying six types of distortions (labeled C2 to C7) to part of the data in the evaluation set, which include distortions from Voice over Internet Protocol (VoIP) and Public Switched Telephone Network (PSTN) environments. The original clean speech data from 19LA is retained in 21LA and is labeled as C1 condition.

In our experiments, the models were trained on the 19LA training set, with the 19LA development set used for threshold tuning. The evaluation was conducted on both 19LA and 21LA evaluation sets.

4.2 Experimental Settings

During the training of the detectors, we employed the Adam optimizer with a base learning rate of 10^{-4}, coupled with cosine annealing for learning rate decay. The batch size was set to 200, and the model was trained for 10 epochs. In the case of our training set including six generation algorithms, a total of 24 detectors were trained in the experiments. For the MDP training process of the detector, $\theta_{p_{train}}^{Train}$ was initialized to 0.5, and p_{train} to compute threshold was set to 99%. The decision trees were implemented using the SKLearn library, with Gini impurity serving as the criterion for splitting nodes.

We compared our BiMDM with four competing systems including AASIST [12], SSL [31], RawFormer-SE [17] and RawBMamba [2]. For AASIST and RawBMamba, the results were obtained with the best pre-trained model that their author provided. And for SSL and RawFormer-SE, we trained the model from scratch under the settings mentioned above without fine-tuning for the pre-trained model and data augmentation.

As our method produces a binary decision for fake speech detection, the performance is measured with standard information retrieval metrics, including *Precision*, *Recall*, $F_1 Score$ and *Accuracy*, higher values of which indicate better performance. The detection results of baselines are obtained with the score thresholds that achieves the best Accuracy.

4.3 Results and Analysis

4.3.1 Final Detection Performance

The final detection performance comparison is conducted from two perspectives including overall detection performance and breakdown detection performance across different subset. The overall detection performance is calculated across the entire speech dataset. Breakdown detection performance, in the 19LA evaluation set, is calculated separately across the thirteen kinds of fake speech generated by different algorithms. In the 21LA evaluation set, it is calculated separately across the seven kinds of signal distortion conditions.

The overall detection performance is presented in Table 1. The evaluation metrics include $F_1 Score$ and $Accuracy(\%)$, both of which are crucial for assessing detection performance. On both 19LA and 21LA datasets, our method outperforms the four competing baseline systems. On the 19LA dataset, we achieves an $F_1 Score$ of 0.9974, which is a notable improvement over AASIST's 0.9884 and SSL's 0.9662. Additionally, our method achieves an accuracy(%) of 99.54, surpassing AASIST (97.95), SSL (94.14), RawFormer-SE(98.80) and RawBMamba(99.13). On the 21LA dataset, the performance gap between our method and the four baseline systems is even larger. Our method achieves an $F_1 Score$ of 0.9827, 0.0192 higher than AASIST, 0.0369 higher than SSL, 0.0036 higher than RawFormer-SE and 0.0041 higher than RawBMamba. And the $Accuracy(\%)$ of our method is 96.84, 1.09 higher than AASIST and 3.07 higher than SSL, 0.66 higher than RawFormer-SE and 0.76 higher than RawBMamba. These results shows the effectiveness of our new methods in solving fake speech detection problems.

Regarding the breakdown detection performance, Figure 4 shows the results in terms of Recall on LA19 evaluation set across the thirteen kinds of fake speech from A07-A19 algorithms. Here, since the prediction for genuine speech remains consistent across different subset performance, we concentrated solely on the value in terms of Recall to assess how well the fake speech from dif-

Table 1: Overall detection performance comparison on 19LA and 21LA evaluation set.

	19LA		21LA	
Method	F_1**Score**	**Accuracy(%)**	F_1**Score**	**Accuracy(%)**
AASIST [12]	0.9884	97.95	0.9635	95.75
SSL [31]	0.9662	94.14	0.9458	93.77
RawFormer-SE [17]	0.9933	98.80	0.9791	96.18
RawBMamba [2]	0.9952	99.13	0.9786	96.08
Our BiMDM	**0.9974**	**99.54**	**0.9827**	**96.84**

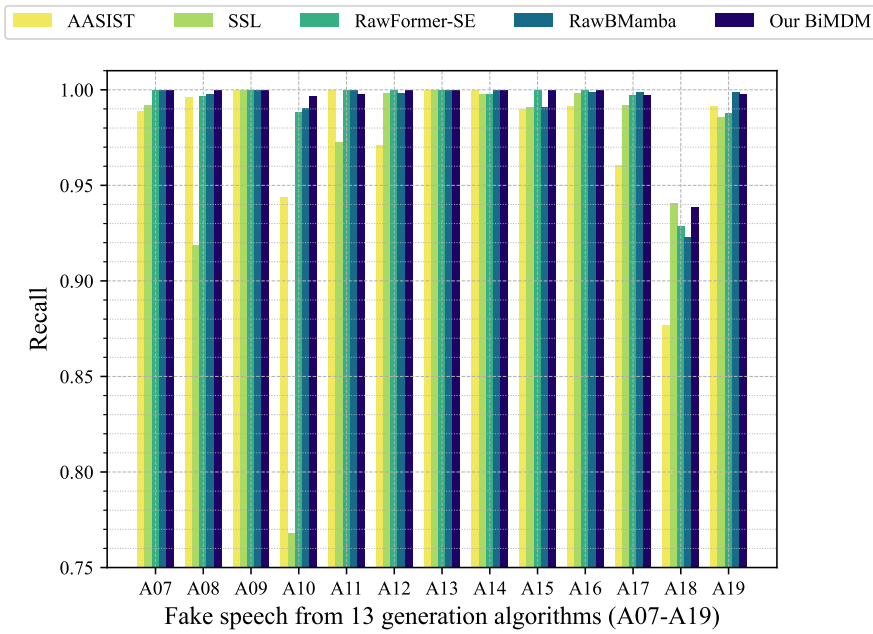


Figure 4: Breakdown detection performance in terms of Recall across fake speech from 13 algorithms (A07-A19) on LA19 evaluation set.

ferent generation algorithms can be detected. Compared with the baseline system AASIST, SSLAug and RawFormer-SE, our BiMDM achieves higher Recall values, nearing or reaching 1.0 for fake speech generated by 12 out of 13 spoofing algorithms, with the exception of A18. The decrease in performance for A18 can be attributed to missed forgery clues, suggesting that further improvements could be made by developing new detectors specifically designed to capture these clues. And the performance of our method is also comparable with RawBMamba, which shows more performance degradation

for A18. In addition, across the 17 subsets, our method obtains more stable performance, which highlights that our method is capable of handling various fake speech type.

The breakdown detection performance on 21LA evaluation set across seven kinds of distortion conditions (C1-C7) is shown in Table 2. Here, the genuine and fake speech are with the same signal distortion in each condition, and we focus on the detection Accuracy(%) to evaluate the model's robustness across different distortion conditions. Although our BiMDM obtains the lowest Accuracy(%) value with 94.16 under C7 condition, the performance in terms of Accuracy(%) exceeds 95.00 under the other conditions. But for the four baseline systems, the Accuracy(%) value less than 95.00 occurs under more than one condition. And compared with the four baseline systems, our BiMDM achieves less performance gap between the original clean speech condition (C1) and the distortion conditions (C2-C7). These results indicate that our BiMDM has superior generalization ability across a variety of signal distortion scenarios.

Table 2: Breakdown detection performance in terms of Accuracy(%) across seven conditions (C1-C7) on 21LA evaluation set.

Method	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
AASIST [12]	97.04	94.89	94.77	95.88	95.75	96.51	93.69
SSL [31]	95.08	92.90	91.93	93.92	93.62	94.29	92.05
RawFormer-SE [17]	98.10	92.80	94.88	96.97	97.52	95.11	92.58
RawBMamba [2]	98.10	96.99	92.19	98.08	97.07	92.82	97.33
Our BiMDM	98.13	97.81	95.42	97.91	97.79	96.62	94.16

4.3.2 Detectors' Output of BiMDM

For the explanation of the detection results, we trace back to the detectors' outputs. If a detector outputs the value of True, it is viewed as activated and the corresponding clue appears in the detection process. Here, for the 13 types of fake speech data in 19LA evaluation set, we count the total times that each of the 24 detectors is activated during the detection process based on different types. The results are shown in Figure 5. For each type of fake speech, several detectors are activated significantly more frequently than others. For fake speech from A07 algorithm, the detectors of SA_4, PA_4 and WA_1 are most frequently activated. For fake speech from A10 algorithm, the detectors of WA_1, and WA_4 are most frequently activated. While for fake speech from A17, A18 and A19 algorithm, the detectors of PA_6 are most frequently activated. This aligns with our common sense that different types of fake speech may contain different forgery clues. And by analyzing

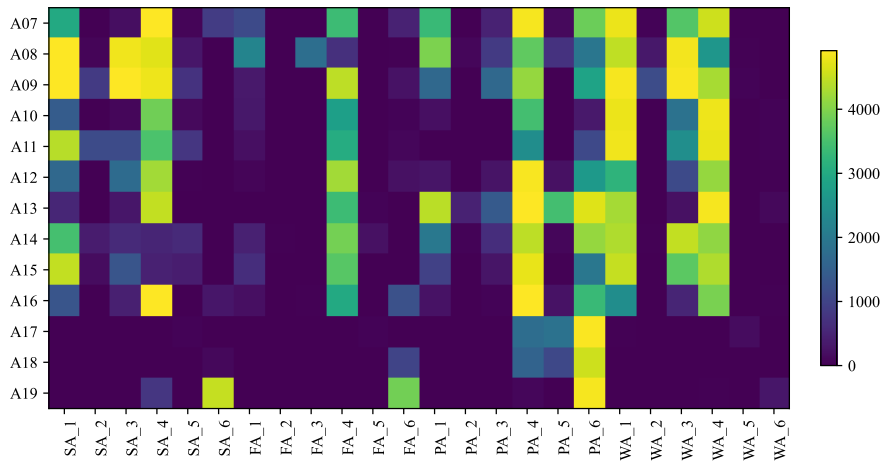


Figure 5: The total times that each of the 24 detectors (SA_1 to SA_6, FA_1 to FA_6, PA_1 to PA_6 and WA_1 to WA_6) is activated during the detection process, according to the 13 types (A07-A19) of fake speech data in the 19LA evaluation set.

which detectors are activated during the detection process, the detection result obtains an explanation of which forgery clue information has been found, leading to this outcome. For example, we trace back the detectors' results for detecting the audio in the name of LA_E_5246322, which is fake speech and generated through algorithm A07. It is detected as fake speech in the final result. The detectors of SA_4 and PA_4 is activated while the others are not activated. Then the explanation of detection result with fake speech class for this audio file is that the clues of SA_4 and PA_4 are discovered, which exhibit characteristics similar to those of fake speech generated by the A04 algorithm in both the spectrogram and phoneme domains, and such discovery is assumed to be valid according to the decision tree "if-else" reasoning.

In addition, in the Figure 5, we found some detectors like FA_2, WA_5 and WA_6 were least frequently activated. This situation does not indicate that these detectors are useless and can be discarded. Since the results are obtained on the 19LA evaluation set which does not include all the fake speech in the world, the above mentioned detectors can also be retained for further use.

4.3.3 Ablation Study

To assess the impact of the detection precision penalty term, we conduct the ablation study with different λ values of penalty term. Here, the case that $\lambda = 0$ means that no detection precision penalty is added. Table 3

Table 3: Overall detection performance on 19LA and 21LA evaluation sets with different penalty term λ. The case of $\lambda = 0$ means no detection precision penalty.

	19LA				21LA			
λ	**Precision**	**Recall**	F_1**Score**	**Accuracy**	**Precision**	**Recall**	F_1**Score**	**Accuracy**
0	0.9994	0.9910	0.9952	0.9914	0.9700	0.9950	0.9823	0.9678
1	0.9994	0.9937	0.9964	0.9936	0.9686	0.9966	0.9824	0.9680
2	0.9994	0.9943	0.9968	0.9943	0.9686	0.9970	0.9827	0.9684

Table 4: Detection performance on 19LA and 21LA evaluation sets with different decision making strategies. The case of $\lambda = 0$ means no detection precision penalty.

	19LA				21LA			
Decision Type	**Precision**	**Recall**	F_1**Score**	**Accuracy**	**Precision**	**Recall**	F_1**Score**	**Accuracy**
OR	0.9990	0.9945	0.9967	0.9942	0.9481	0.9995	0.9732	0.9504
w/ decision trees	0.9994	0.9943	0.9968	0.9943	0.9686	0.9970	0.9827	0.9684

presents the overall detection performance on the 19LA and 21LA evaluation sets. On the 19LA dataset, we observe an improvement of overall detection performance in terms of F_1Score and *Accuracy* with the increasing λ from 0 to 2. F_1Score increases from 0.9952 to 0.9968 and *Accuracy* from 0.9914 to 0.9943. But on the 21LA dataset, such increase is minor with F_1Score from 0.9823 to 0.9827 and *Accuracy* from 0.9678 to 0.9684. As there are seven distortion conditions on the audios in the 21LA evaluation set, we also make a breakdown detection performance comparison across these conditions. The detail performance comparison in terms of Precision and Accuracy is shown in Figure 6. In most conditions, the performance with and without the penalty term is close and remains comparable to the original condition, except for C3 and C7. There are performance degradations under the two conditions, which can be attributed to the increased degree of distortion. Even so, the method with the detection penalty term (i.e. $\lambda = 1$ and $\lambda = 2$) shows less degradation than that without the penalty term (i.e. $\lambda = 0$). This demonstrates that the extra detection penalty term can enhance the detection robustness for different audio transformation distortions.

For decision-making process, we compare the strategies of only OR logic calculus and decision trees with OR logic. As the results shown in Table 4, for 19LA dataset, the difference between the two decision-making strategies is minimal. But for 21LA dataset, the strategy of decision trees shows clear improvements with F_1Score (0.9732 to 0.9827) and *Accuracy* (0.9504 to 0.9684). A simple OR logic to combine all detector outputs will cause an accumulation of errors when the detectors make mistakes, where the mistakes on 19LA data are not obvious, but they are more on 21LA data with signal distortion. On 21LA dataset, the strategy of decision trees achieves an improvement of 0.0205 in terms of Precision value, which demonstrates its ability of error tolerance to detectors.

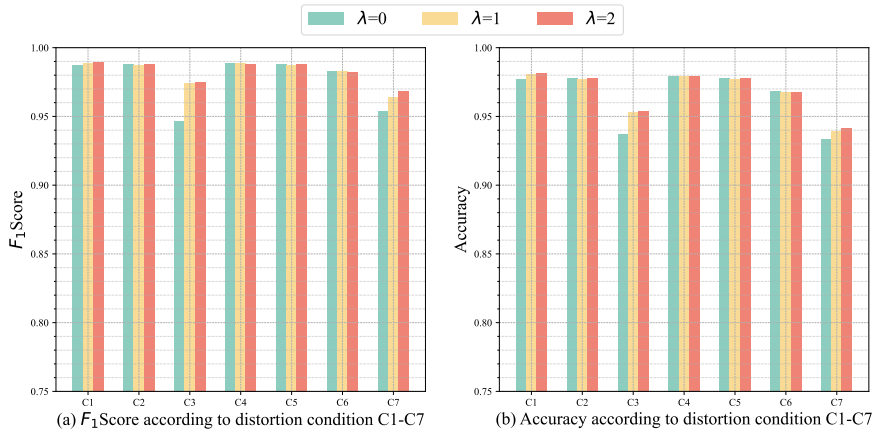


Figure 6: For different penalty term λ, the breakdown detection performance in terms of (a) Recall and (b) Accuracy, across seven distortion conditions (C1-C7) on 21LA evaluation set.

5 Conclusion

This paper proposed a Brain-inspired Multi-Detector Machine (BiMDM) for fake speech detection task. It contains multiple independent detectors, each specialized in detecting specific forgery clues with the aim of Maximum Detection Precision (MDP), moving beyond the traditional classification perspective. Then the detectors' binarazed results of bool type are integrated by the decision-making module built upon Decision Trees and OR logic calculus. Our experimental results showed the superior detection performance of our BiMDM, comparing with the baselines. And the breakdown detector performance according to different subsets demonstrates that our method is robust across various fake speech types as well as speech signal distortion conditions. Additionally, we traced the detection process back to the individual detector outputs, highlighting the potential for providing explanations of the final detection result.

Overall, the proposed method offers a novel perspective on fake speech detection by constructing global detection information from multiple individual fake speech forgery clues. And the decision-making module of our BiMDM is a sequential structure to integrate detector outputs, enabling the potential for sequential learning of new knowledge. Our future work will focus on exploring BiMDMs adaptability to new knowledge, in order to face the continuous advancement of fake speech generation techniques and new attack types.

References

[1] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale", in *Proc. of 2022 Interspeech*, Incheon: International Speech Communication Association, 2022, 2278–82.

[2] Y. Chen, J. Yi, J. Xue, C. Wang, X. Zhang, S. Dong, S. Zeng, J. Tao, Z. Lv, and C. Fan, "RawBMamba: End-to-End Bidirectional State Space Model for Audio Deepfake Detection", in *Proc. of 2024 Interspeech*, Kos: International Speech Communication Association, 2024, 2720–4.

[3] C. Du, Y. Guo, X. Chen, and K. Yu, "Speaker Adaptive Text-to-Speech With Timbre-Normalized Vector-Quantized Feature", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2023, 3446–56.

[4] C. Fan, J. Xue, S. Dong, M. Ding, J. Yi, J. Li, and Z. Lv, "Subband Fusion of Complex Spectrogram for Fake Speech Detection", *Speech Communication*, 155, 2022, 102988.

[5] C. Feng, Y. Zhao, G. Sun, Z. Chen, S. Wang, C. Zhang, M. Xu, and T. F. Zheng, "Hierarchical Multi-Path and Multi-Model Selection For Fake Speech Detection", in *Proc. of 2024 IEEE Spoken Language Technology Workshop (SLT)*, Macau: IEEE, 2024, 983–90.

[6] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj, and R. Singh, "Generalized Spoofing Detection Inspired from Audio Generation Artifacts", in *Proc. of 2021 Interspeech*, Brno: International Speech Communication Association, 2021, 4184–8.

[7] W. Ge, M. Panariello, J. Patino, M. Todisco, and N. Evans, "Partially-Connected Differentiable Architecture Search for Deepfake and Spoofing Detection", in *Proc. of 2021 Interspeech*, Brno: International Speech Communication Association, 2021, 4319–23.

[8] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection", in *Proc. of 2019 Interspeech*, Graz: International Speech Communication Association, 2019, 1068–72.

[9] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable Text-To-Speech With Text Descriptions", in *Proc. of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island: IEEE, 2023, 1–5.

[10] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for Synthetic Speech Detection: A Comparison", in *Proc. of 2015 Interspeech*, Dresden: International Speech Communication Association, 2015, 2057–61.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *Proc. of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, 2016, 770–8.

[12] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks", in *Proc. of 2022 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, Singapore, 2022, 6367–71.

[13] C. Kirchhübel and G. Brown, "Spoofed speech from the perspective of a forensic phonetician", in *Proc. of 2022 Interspeech*, Incheon, 2022, 1308–12.

[14] D. Kumar, P. K. V. Patil, A. Agarwal, and S. M. Prasanna, "Fake Speech Detection Using OpenSMILE Features", in *Proc. of the 24th International Conference on Speech and Computer*, ed. S. R. M. Prasanna, A. Karpov, K. Samudravijaya, and S. S. Agrawal, Gurugram: Springer, 2022, 404–15.

[15] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and Synthetic Speech Detection with Res2Net Architecture", in *Proc. of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto: IEEE, 2021, 6354–8.

[16] R. Liu, B. Sisman, G. Gao, and H. Li, "Controllable Accented Text-to-Speech Synthesis With Fine and Coarse-Grained Intensity Rendering", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 2024, 2188–201.

[17] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging Positional-Related Local-Global Dependency for Synthetic Speech Detection", in *Proc. of 2023 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island: IEEE, 2023, 1–5.

[18] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2023, 2507–22.

[19] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Neutral-to-Emotional Voice Conversion with Cross-Wavelet Transform F0 Using Generative Adversarial Networks", *APSIPA Transactions on Signal and Information Processing*, 8(1), 2019.

[20] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake Audio Detection Based on Unsupervised Pretraining Models", in *Proc. of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore: IEEE, 2022, 9231–5.

[21] M. Pal, D. Paul, and G. Saha, "Synthetic Speech Detection Using Fundamental Frequency Variation and Spectral Features", *Computer Speech & Language*, 48, 2018, 31–50.

[22] J. Pesnot Lerousseau, C. V. Parise, M. O. Ernst, and V. van Wassenhove, "Multisensory Correlation Computations in the Human Brain Identified by A Time-Resolved Encoding Model", *Nature communications*, 13(1), 2022, 2489.

[23] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with Sincnet", in *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Athens, 2018, 1021–8.

[24] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A Comparison of Features for Synthetic Speech Detection", in *Proc. of 2015 Interspeech*, Dresden: International Speech Communication Association, 2015, 2087–91.

[25] J. Sanchez, I. Saratxaga, I. Hernáez, E. Navas, D. Erro, and T. Raitio, "Toward A Universal Synthetic Speech Spoofing Detection Using Phase Information", *IEEE Transactions on Information Forensics and Security*, 10(4), 2015, 810–20.

[26] A. K. Singh and P. Singh, "Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics", in *Proc. of the 4th International Conference on Multimedia Information Processing and Retrieval*, Tokyo: IEEE, 2021, 412–7.

[27] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of Sub-Band Discriminative Information between Spoofed and Genuine Speech", in *Proc. of 2016 Interspeech*, San Francisco: International Speech Communication Association, 2016, 1710–4.

[28] H. Tak, J.-w. Jung, J. Patino, M. Todisco, and N. Evans, "Graph Attention Networks for Anti-Spoofing", in *Proc. of 2021 Interspeech*, Brno: International Speech Communication Association, 2021, 2356–60.

[29] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing Attack Detection Using the Non-Linear Fusion of Sub-Band Classifiers", in *Proc. of 2020 Interspeech*, ISCA, Shanghai, 2020, 1106–10.

[30] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End anti-spoofing with RawNet2", in *Proc. of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto: IEEE, 2021, 6369–73.

[31] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation", in *Proc. of the 12th Speaker and Language Recognition Workshop Odyssey*, Beijing, 2022, 112–9.

[32] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]", https://doi.org/10.7488/ds/1994, 2017.

[33] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, R. Fu, and X. Chen, "TO-Rawnet: Improving RawNet with TCN and Orthogonal Regularization for Fake Audio Detection", in *Proc. of Interspeech 2023*, International Speech Communication Association, 2023, 3137–41.

[34] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, *et al.*, "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech", *Computer Speech & Language*, 64, 2020, 101114.

[35] Y. Wen, Z. Lei, Y. Yang, C. Liu, and M. Ma, "Multi-Path GMM-MobileNet Based on Attack Algorithms and Codecs for Synthetic Speech and Deepfake Detection", in *Proc. of 2022 Interspeech*, Incheon: International Speech Communication Association, 2022, 4795–9.

[36] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks", in *Proc. of 2020 Interspeech*, Shanghai: International Speech Communication Association, 2020, 1101–5.

[37] Z. Wu, E. S. Chng, and L. Haizhou, "Detecting Converted Speech and Natural Speech for Anti-Spoofing Attack in Speaker Recognition", in *Proce. of the 13th Annual Conference of the International Speech Communication Association*, Portland: International Speech Communication Association, 2012, 1700–3.

[38] Y. Xie, Z. Zhang, and Y. Yang, "Siamese Network with Wav2vec Feature for Spoofing Speech Detection", in *Proc. of 2021 Interspeech*, Brno: International Speech Communication Association, 2021, 4269–73.

[39] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection", in *Proc. of 2023 Interspeech*, Dublin: International Speech Communication Association, 2023, 2808–12.