Scientific Reports

https://doi.org/10.1038/s41598-026-41859-8

Article in Press

Brain-inspired perception-decision machine for fake speech detection

Received: 1 July 2025

Accepted: 23 February 2026

Published online: 05 March 2026

Cite this article as: Feng C., Wu X., Askar H. *et al.* Brain-inspired perception-decision machine for fake speech detection. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-41859-8

Chang Feng, Xiaolong Wu, Hamdulla Askar, Mingxing Xu, Lihong Cao & Thomas Fang Zheng

Brain-inspired Perception-Decision Machine for Fake Speech Detection

Chang Feng[1], Xiaolong Wu[2], Hamdulla Askar[2], Mingxing Xu[1], Lihong Cao[3], and Thomas Fang Zheng[1,*]

[1]Center for Speech and Language Technologies, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, China
[2]School of Computer Science and Technology, Xinjiang University, Urumqi, 830000, China
[3]Neuroscience and Intelligent Media Institute, Communication University of China, Beijing, 100024, China
[*]Corresponding author: fzheng@tsinghua.edu.cn

ABSTRACT

The rapid advancement of Artificial Intelligence Generated Content (AIGC) technologies challenges fake speech detection with an ever-evolving diversity of spoofed audio. Current approaches, which rely on a classification-based perspective, are highly dependent on a big amount of training data and show limited generalization to unseen attack types. To address these limitations, this paper introduces a brain-inspired, multi-clue detection paradigm. We propose a perception-decision machine composed of two core components. The perception module utilizes multiple independent detectors, each optimized for Maximum Detection Precision (MaxDP) to identify a specific forgery clue. By standardizing their outputs into binary Boolean values, this design allows for flexible computational models. The decision-making module then renders a final judgment by first evaluating learned combinations of the detected clues through a logical reasoning process. The outcomes of this reasoning are then aggregated using a variable-length OR operation, a mechanism that enables the seamless incremental learning of new forgery clues without retraining the entire system. Our results validate the effectiveness of the multi-clue detection perspective, demonstrating the framework's potential for enhanced explainability and practical adaptability to new threats.

Introduction

The advancement of Artificial Intelligence Generation Content (AIGC) technologies have made it easy to generate fake speech audios that closely mimic genuine human voices. The growing accessibility to fake speech audios introduces serious threats to speech-based applications, including voice authentication, voice-controlled services and voice communication systems. Attackers can exploit fake speech to impersonate individuals, deceive automated systems, or disseminate disinformation, leading to privacy leakage, financial fraud, and user trust erosion. These growing threats highlight the need for fake speech detection. One of the key challenges in fake speech detection lies in the diversity of fake speech, which results from the diversification of speech generation algorithms with text-to-speech synthesis (TTS)[1] and voice conversion (VC)[2]. There are many observations about the differences between fake and genuine speech. For example, fake speech may have blurred pitch source harmonics while genuine speech has sharp harmonic structure[3], which can be observed in the low-frequency region of spectrogram. Some fake speech can be characterised by a lack of definition in formant frequencies and other unusual striations in the spectrogram[4]. In addition, different vocoders used for fake speech generation may produce different artifacts[5], and these artifacts tend to appear in different sub-bands on the spectrogram[6,7]. The existing solutions view fake speech detection task as a classification task. Most approaches focus on enhancing the discriminative power of classifiers through improved feature extraction techniques[8–10] and effective model architectures[11–13]. In an attempt to further improve robustness against diverse types of fake speech, some studies adopt ensemble learning strategy[14–16] which combines outputs from multiple base classifiers to mitigate the limitations of the individual classifier. This strategy typically computes a weighted sum of prediction scores from base classifiers, using either equal (average) or unequal (non-average) weights. However, the existing fake speech detection solutions, which are predominantly based on classification, face three critical limitations that hinder their real-world applicability. First, they take the classification-based perspective[17] and aim at Minimum Classification Error (MinCE) by automatically refining decision boundaries from examples, making their performance highly dependent on large-scale training dataset. As the diversity of fake speech increases, the complexity of modeling all the data increases and so the detection performance will decrease. Second, they exhibit poor generalization to unseen data. These computation models typically treat all types of fake speech uniformly during training without modeling their specific traits, their performance degrades significantly when confronted with unseen forgery algorithms not present in the training data. Finally, they lack the ability of incremental learning for new forgery clues and necessitate costly retraining when new knowledge or data is added. The "black-box" nature

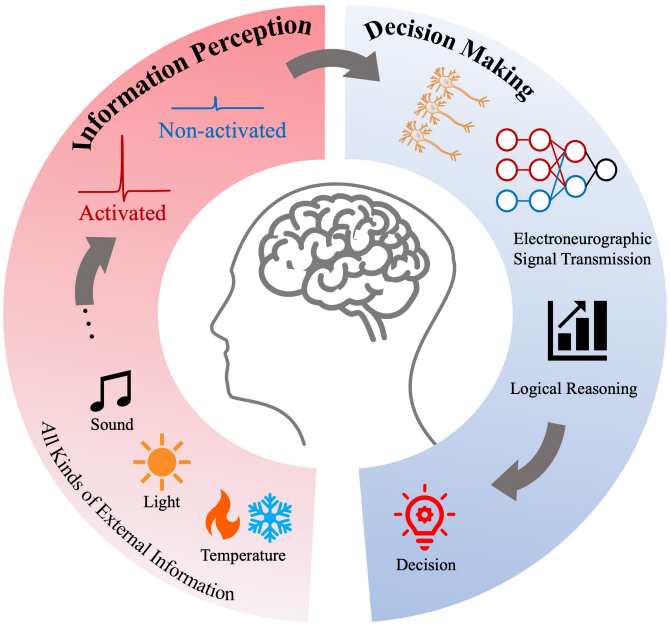


Figure 1. In biological systems, brain activity process for the coming external information includes information perception and decision making.

of classifier models makes it difficult to incrementally incorporate new knowledge of forgery attributes. When a new type of fake speech emerges, developers cannot simply update the model but must instead undertake a full retraining process, which is both inefficient and computationally expensive.

To overcome these limitations, it is necessary to explore new methodologies that go beyond conventional classification-based frameworks and better align with human cognitive processes. Humans typically detect fake speech by identifying unnatural clues that are implicitly preset based on prior knowledge and experiences. In the auditory assessments[4], for example, human experts assess speech audio authenticity by identifying whether a series of unnatural clues appear, such as abnormal phonation of certain phonemes[18], or unusual prosodic features including variations in pitch, loudness, duration, and intonation[19]. Such process is analogous to the way that the brain makes decisions on the basis of perceived information (as in Figure 1), from multiple specialized sensory pathways[20]. In biological systems, external information with multiple modality is perceived and converted into electroneurographic signal through sensory receptors with different structures and mechanisms. Here, each receptor is preconfigured to detect a particular type of information. For example, external temperature information is perceived by thermoreceptors which include two types of receptors specifically responsible for detecting cold or warm stimuli. Also, the external light information is perceived by photoreceptors which include different types of receptor cells, each detecting specific wavelengths of light. Structurally, these receptors are independent and modular, allowing partial updates or replacements without affecting the others. Functionally, these receptors act as abstraction units that extract heterogeneous external inputs into standardized electroneurographic signals in a binary fashion—either activated or non-activated—depending on whether its specific information is detected as present. After information perception, the electroneurographic signals with perceived information are transmitted to modality-specific brain regions for further feature extraction process and multisensory integration for decision-making process[21,22]. This decision-making process can be interpreted by a series of logical rules and conditional judgments.

Inspired by the above-mentioned process in the brain, we introduce a new perspective based on multi-clue detection for fake speech detection task. As there are many forgery clues in fake speech audios, the fake speech detection problem can be broken down into detecting multiple forgery clues, reducing the difficulty of each detection computation and increasing the transparency of the detection process. This process is analogous to the high-level cognitive mechanism used by the human brain to process complex stimuli. First, in the perception phase, specialized neural populations are serve to perceive independent cognitive clues. Second, these clues are transmitted to higher-order association areas, where they are integrated through logical reasoning and conditional judgments to form a final cognition. We propose a brain-inspired perception-decision machine with perception module and decision-making module. The perception module contains multiple detectors for specific forgery clues that are independent with each other. Different from previous classifiers aimed at MinCE, these detectors are defined with the aim of Maximum Detection Precision (MaxDP) of 100%. Theoretically, we posit that there exists a foundational

set of detectors that are minimum coverage, representing the theoretical lower bound on the number of detectors required to perceive the complete set of forgery clues. This minimal set ensures that, in principle, all fake speech can be identified. Each detector consists of a computational model and a threshold that is set to maximize the precision ratio. By comparing the model score with the threshold, the detector output is standardized as a binary Boolean value to indicate whether the specific clue is present or not. Such standarization process allows for flexibility in detector's computational model design, that is, traditional modeling approaches and neural network methods are practicable, which can utilize computaional ability as much as possible. Then the detector outputs are input into the decision-making module to make the final judgment. In theory, the simplest decision-making strategy can be modeled as an OR logic operation, where an audio will be deemed as fake speech if any single forgery clue is present. And to further enhance accuracy, the decision-making process should jointly consider multiple perceived clues, which is manifested as a logical reasoning procedure with evaluations of the co-occurrence and consistency among the clues. And such logical reasoning is modeled with OR logic operation connecting multiple decision trees which are constructed through reinforcement learning with information gain as reward metrics. Furthermore, a significant advantage of our proposed architecture is its inherent support for incremental learning, enabling the system to adapt to a continuously evolving threat landscape. Given that new forgery algorithms are constantly under development, the corresponding forgery clues are also in a perpetual state of flux, necessitating that any practical detection system must be able to iteratively update and incrementally learn these emerging features. As novel forgery techniques emerge, new detectors can be independently designed and trained to identify these previously unknown clues. These newly developed detectors, adhering to the same design principles, can be seamlessly integrated into the perception module. Their binary outputs are then incorporated as new inputs into the decision-making module, effectively expanding the system's knowledge base without the need to retrain the entire model. Consequently, the system can continuously augment its defensive capabilities over time, ensuring robust performance against future threats.

Below, the experiments are conducted on ASVspoof2019[23] Logical Access (LA) and the ASVspoof2021[24] LA datasets. The results demonstrate that fake speech detection problem with diverse forgery characters can be solved from the multi-clue detection perspective. And the detection process can be explained and accessible from a human perspective. Furthermore, our perception-decision machine shown to be flexible to incrementally learn new knowledge of forgery attributes through the addition of new detectors, with experiments on Chinese Dataset for Fake Audio Detection (CFAD)[25].

Results

Dataset

We conducted the detection performance experiments on ASVspoof2019[23] Logical Access (19LA) and the ASVspoof2021[24] LA (21LA) datasets. The 19LA dataset consists of genuine speech alongside fake speech from nineteen generation algorithms of speech synthesis and voice conversion techniques, under consistent channel conditions. It is divided into three mutually exclusive subsets: the training set, which is used during the training phase; the development set, used for tuning; and the evaluation set, used for result validation. The training and development sets contain fake speech generated by six algorithms (A01–A06), while the evaluation set comprises fake speech from the other thirteen unseen algorithm types (A07–A19), providing a more rigorous test of generalization capability. The 21LA dataset, designed to verify the robustness of models, contains only an evaluation set. The speech samples are generated by applying six types of channel transmission distortions (C02–C07) to the audios from the evaluation set of the 19LA dataset.

To verify the flexibility to incrementally learn new knowledge of forgery attributes, we expanded our model to Chinese Dataset for Fake Audio Detection (CFAD) dataset[25], using its clean version. Following the official protocol, the training and development sets include audio from eight spoofing algorithms (F01–F08). The evaluation set contains both a seen scenario with the same eight algorithms and an unseen scenario with four entirely new algorithms (F09–F12).

Comparing detection performance

To validate the effectiveness and robustness of our proposed method, we conducted a comparative analysis against four prominent baseline systems:

- Baseline1: AASIST[26], an end-to-end audio anti-spoofing system based on graph neural networks. It models temporal and spectral features through a heterogeneous stacking graph attention layer (HS-GAL) with a stack node and integrates information via a novel max graph operation (MGO).

- Baseline2: SSL[27], a system applies a fine-tuned wav2vec 2.0 model as a self-supervised front-end to extract robust speech representations, combined with a spectro-temporal graph attention back-end (AASIST) for spoofing detection. It further incorporates a self-attentive aggregation layer and data augmentation.

- Baseline3: RawFormer-SE[28], a system leverages positional-related local-global dependencies by combining a RawNet2 front-end with a Transformer-based classifier. Its positional aggregator is to preserve spectro-temporal information.

	Development Set(Seen Types)		Evaluation Set(Unseen Types)	
Method	F_1**score**	**Accuracy**	F_1**score**	**Accuracy**
Baseline1	0.9989	0.9981	0.9884	0.9795
Baseline2	0.9998	0.9998	0.9662	0.9414
Baseline3	0.9998	0.9996	0.9933	0.9880
Our Method	0.9998	0.9998	**0.9975**	**0.9954**

Table 1. Overall detection performance on ASVspoof2019 LA dataset. Performance comparison with the baseline approaches on the development set (seen types) and evaluation set (unseen types). The results of baselines are obtained by setting the score threshold that makes the best Accuracy.

The overall detection performance of our method compared to three baselines is summarized in Table 1. On the development set (seen types), all models demonstrated a ceiling effect, achieving nearly perfect and highly comparable results with both F_1score and Accuracy. However, the transition to the evaluation set (unseen types) created a clear performance divergence. The baseline models exhibited a notable degradation, exposing their limited ability to generalize. This performance gap was particularly pronounced for models like Baseline2. Even the strongest competitor, Baseline3, while performing well, still showed a discernible drop in both F_1score and Accuracy when faced with unseen generation algorithm types. In contrast, our proposed method showcased superior robustness by maintaining exceptional performance with only a minimal decline. By achieving a state-of-the-art F_1score of 0.9975 and an Accuracy of 0.9954 on the unseen types, our method demonstrated a significantly smaller generalization gap than any baseline. This direct comparison validates that while all models appear competent on familiar data, our method's architecture is uniquely effective at abstracting fundamental forgery principles, enabling it to remain highly accurate and robust against new, previously unseen forgery techniques.

The evaluations were performed on the ASVspoof 2019 (19LA) and 2021 (21LA) datasets, with the comprehensive results presented in Figure 2. First, we assessed the generalization capability of the models against 13 unseen fake speech algorithms (A07–A19) from the 19LA dataset. Since the model demonstrated a stable true negative rate (TNR) on genuine speech across the various test conditions, we selected Recall as the key performance indicator to specifically evaluate its ability to detect fake speech samples. As illustrated in Figure 2(a), our method demonstrated superior and more consistent recall performance across nearly all fake speech algorithms compared to the baselines. Notably, our method achieved a perfect recall of 100% on five of the thirteen algorithms (A07, A09, A13, A14, and A15). In cases where other models struggled, such as Baseline2 against algorithm A10 (76.80% recall) or Baseline1 against A18 (87.68% recall), our model maintained exceptionally high performance. This indicates a significant improvement in generalization to diverse and previously unseen forgery types. Next, the model's robustness is evaluated under the seven different channel-distorted conditions (C1–C7) of the 21LA dataset. The results, shown in terms of F1-score (Figure 2(b)) and Accuracy(%) (Figure 2(c)), reveal a significant and near-uniform advantage for our perception-decision machine method. Specifically, our method achieved the highest F1-score on the seven conditions (C1–C6), signifying a superior balance between precision and recall under most real-world transmission scenarios. This trend was mirrored in the accuracy measurements.

Tracing back to the detection process

To understand the detection process of our model and evaluate its robustness, we analyzed which specific detectors were activated by different types of fake speech under various channel conditions. For each of the 13 unseen spoofing algorithms (A07–A19), we measured the activation frequency of each detector across the seven channel-distorted conditions (C1–C7) of the 21LA dataset. The results are visualized as heatmaps in Figure 1.

A key finding from this analysis is the stability of detector activation patterns against channel distortion, which demonstrates the robustness of our multi-detector approach. As shown across the seven heatmaps in Figure 3 under seven channel-distorted conditions (C1–C7), the set of most frequently activated detectors remains largely consistent regardless of the channel condition. For example, the activation pattern for algorithm A07 (the top row) shows a similar set of prominent detectors being triggered across all conditions. An even more striking example is algorithm A17, which reliably activates specific detectors with high frequency across all seven conditions. This consistency suggests that the underlying forgery clues captured by our detectors are fundamental to the spoofing algorithm itself and are not easily masked by channel effects.

The activation of a detector signifies that a specific, predefined forgery clue has been found. By tracing back to the detectors' outputs, we can interpret why an audio sample was classified as fake. For instance, a sample generated by algorithm A19 is flagged as fake primarily because it contains the clue associated with the consistently activated detector on the far right.

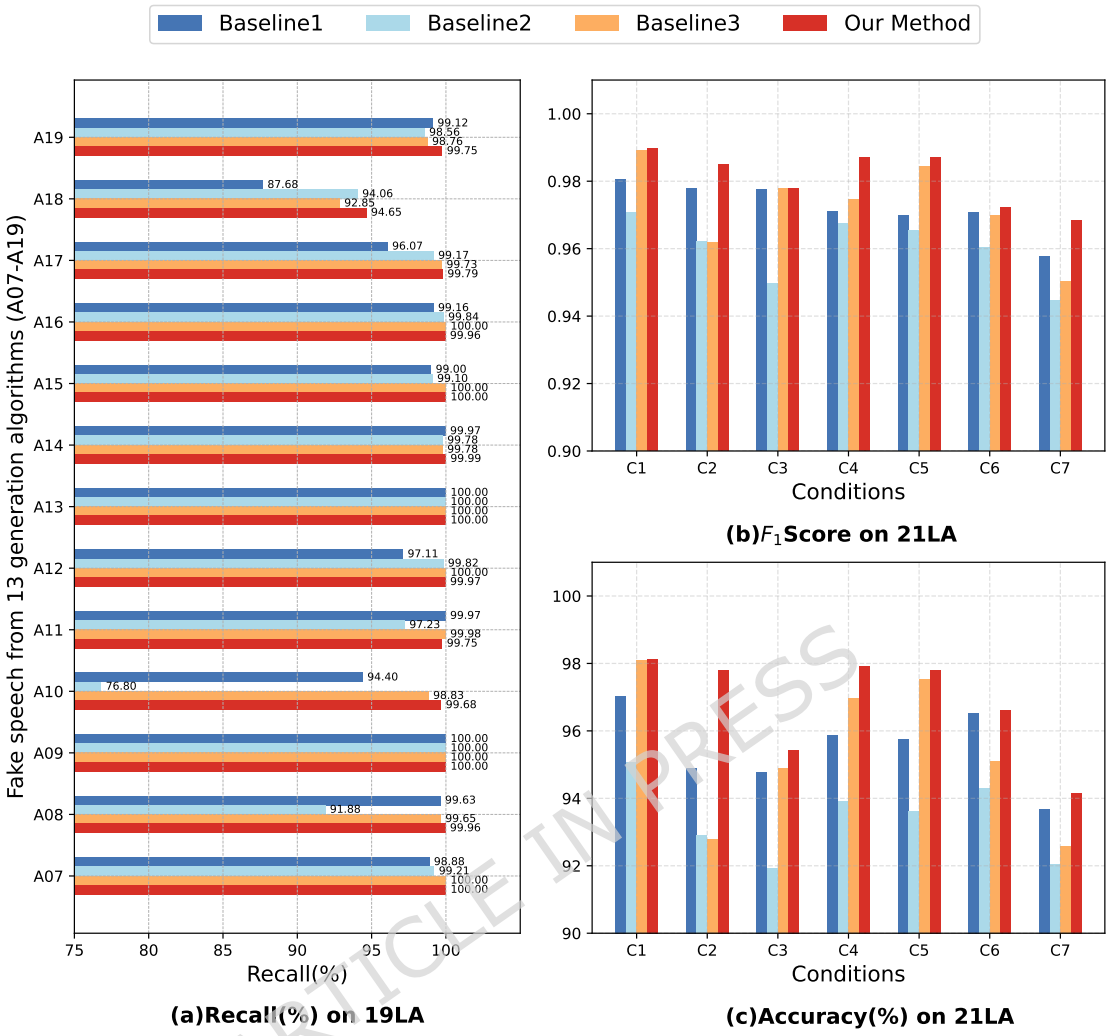


Figure 2. The detection performance comparison on the evaluation set of 19LA and 21LA for each subset.

In contrast, a sample from algorithm A10 is identified due to a different set of activated detectors. This ability to attribute a detection result to a specific set of discovered artifacts provides a clear and meaningful explanation, moving beyond the black-box nature of conventional classifiers.

Finally, we note that some detectors appear to be infrequently activated across this specific set of 13 algorithms (visible as consistently dark vertical bands). This does not indicate that these detectors are redundant. Since the evaluation set cannot encompass every possible generation technique in existence, these detectors are retained within the framework. They may be crucial for identifying forgery clues present in novel or future attack types, ensuring the system's forward compatibility and broad-coverage potential.

Incrementally learning for new clue knowledge

A key advantage of our perception-decision machine is its designed flexibility and capacity for incremental learning. To validate this, we conducted an experiment to augment the existing system with new detectors tailored for new forgery clues. The modular nature of the detectors, combined with the variable-length OR logic in the decision-making module, allows for new capabilities to be added without retraining the existing detectors.

Initially, all models were trained on the 19LA. Subsequently, they were updated to incorporate new knowledge from CFAD. For the three baseline systems, this update was performed via full re-training on a combined dataset. For our method, we simply trained new, specialized clue detectors for the CFAD data and added it to our existing perception module. The results, presented

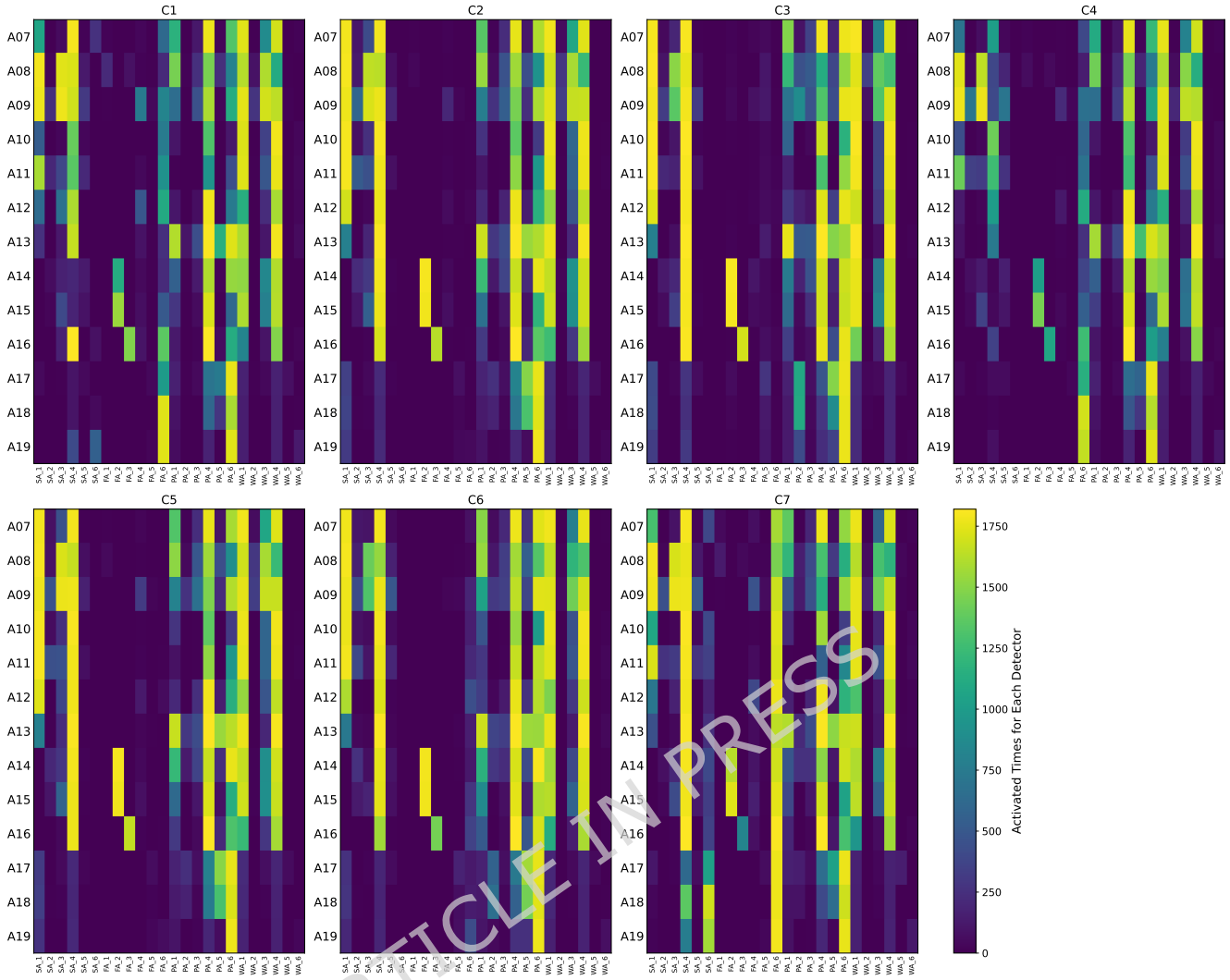


Figure 3. Statistical data on the activation times for each of the 24 detectors (SA_1 to SA_6, FA_1 to FA_6, PA_1 to PA_6, and WA_1 to WA_6) recorded during the detection process categorized by the 13 generation algorithms (A07–A19) of fake speech data in the 21LA evaluation set, under seven channel-distorted conditions (C1–C7).

in Figure 4, evaluate the performance of incremental learning for our perception-decision machine.

The baseline methods, as shown in Figure 4(a-c), demonstrate a critical flaw of the re-training paradigm. While re-training successfully enables the models to learn the new CFAD task (achieving F_1score around 0.95), it comes at a great cost to previously acquired knowledge. Catastrophic forgetting is observed, where the models' performance on the original 19LA and 21LA datasets degrades. For instance, after re-training, Baseline1's F_1score on 19LA dropped drastically from 0.9884 to 0.6491, and Baseline3's score on 21LA fell from 0.9791 to 0.5868. This approach is not only computationally expensive and resource-intensive due to the need for complete model re-training, but it also proves unsustainable for incremental learning as it effectively "forgets" how to perform old tasks. In contrast, our proposed method, shown in Figure 4(d), also learned the new CFAD task effectively by adding new forgery clue detectors, achieving a high F_1score of 0.9549. Crucially, it did so with minimal impact on existing knowledge. The performance on the original 19LA and 21LA datasets remained exceptionally high, with F_1score only dropping negligibly from 0.9974 to 0.9746 and from 0.9927 to 0.9698, respectively.

These results demonstrate the superiority of our modular, detector-addition approach for incremental learning. It can efficiently incorporate new knowledge while preserving existing capabilities, avoiding the resource-intensive and destructive cycle of re-training. This makes our framework a far more practical and scalable solution for real-world applications where detection systems must constantly adapt to new and evolving threats.

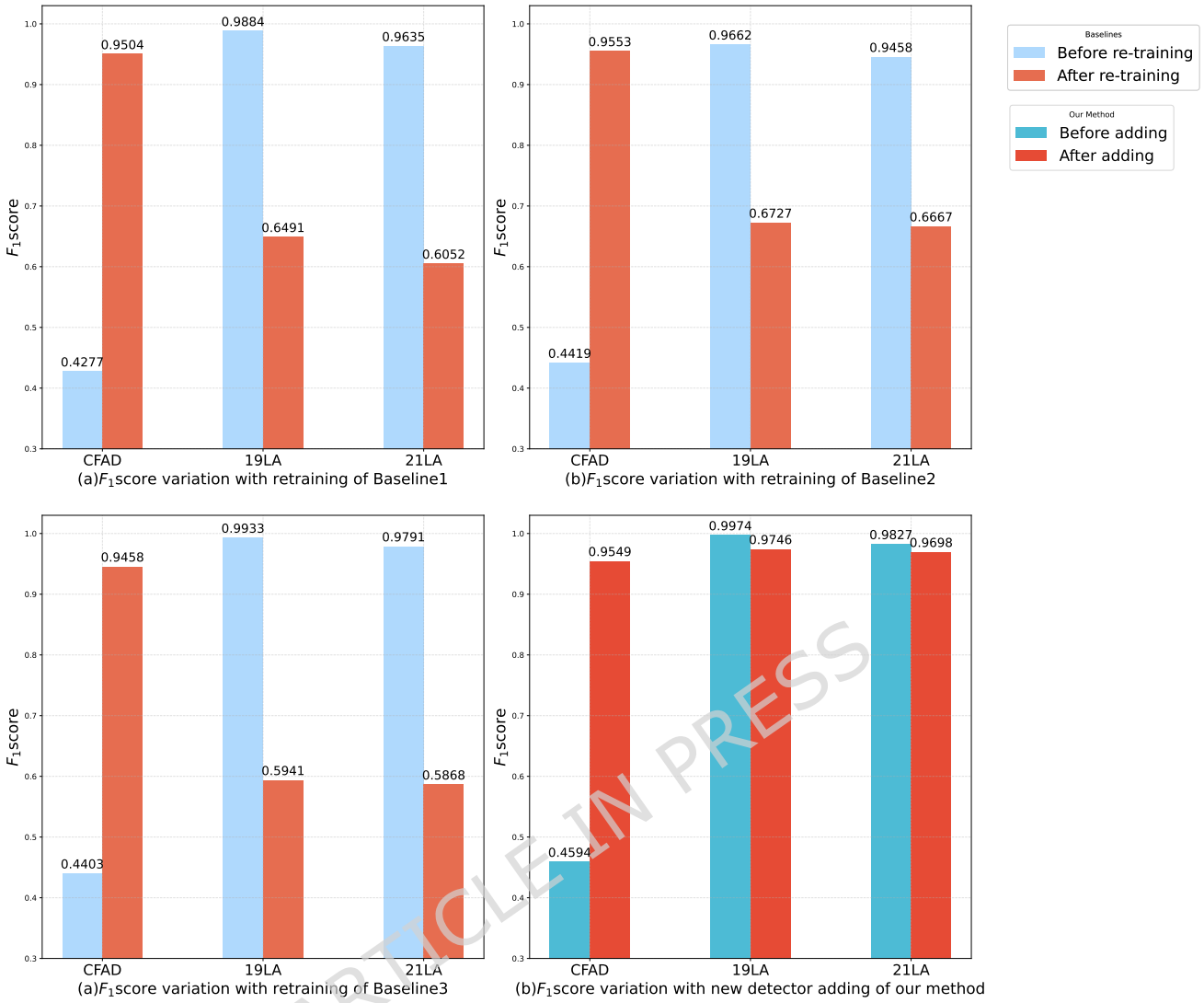


Figure 4. The performance variation for learning new knowledge from new fake speech with chinese language. For the three baseline systems, the strategy is re-training. For our method, new clue detectors are additionally trained and added to the framework.

Discussion

In this study, we introduced a brain-inspired perception-decision machine that reframes the fake speech detection problem from a conventional classification task to a multi-clue detection paradigm. Traditional methods, which aim to minimize classification error (MinCE), struggle with the increasing diversity of speech synthesis artifacts and lack adaptability. Our approach, inspired by the human brain's perception processing, decomposes the problem into detecting multiple, specific forgery clues. Distinct from recent brain-inspired systems that rely on neuromorphic hardware implementations or implicit neural dynamics for adaptive perception[29–31], our framework focuses on functional cognitive mimicry at the algorithmic level. We abstract the biological principles into modular software detectors and explicit logical reasoning, prioritizing the transparency of the decision process over the emulation of physical neural substrates.This modular strategy enhances robustness against diverse fake speech types and provides a transparent, explainable framework.

A key innovation of our work is the design of individual clue detectors that aim for Maximum Detection Precision (MaxDP) rather than Minimum Classification Error (MinCE). By setting a high-precision threshold for each detector, the detector gets a binarized output, where the positive signal of the detector is assumed highly reliable, even if it means some instances with fake label are missed by a single detector. This design allows for the flexible integration of various computational models, from traditional models (e.g. Gaussian mixture models) to deep neural networks, to make effective use of computing power. As

demonstrated by our experiments on the ASVspoof datasets, this collective approach of combining multiple high-precision detectors effectively minimizes the overall missing detection rate and achieves high accuracy. The results confirm that the multi-clue detection paradigm is a viable and effective solution for the fake speech detection challenge. And the introduction of detection threshold provides direct human control over the detector's sensitivity. This stands in stark contrast to conventional black-box methods, where the decision boundary is opaque and does not offer a straightforward mechanism for operator adjustment.

The second component of our machine, the decision-making module, leverages the binary outputs of the clue detectors to make a final judgment. By organizing the detected clues using a combination of OR logic and decision trees constructed via reinforcement learning, this module mimics human-like logical reasoning. This structure not only enhances detection accuracy by considering the co-occurrence of clues but also imbues the entire system with explainability—a critical feature absent in most "black-box" classifiers. For instance, the system can report which specific artifacts led to its decision. Furthermore, the modularity facilitated by the OR logic operation enables straightforward incremental learning. As shown in our experiments with the CFAD dataset, new detectors for novel forgery clues (such as those specific to a different language) can be seamlessly added without retraining the entire system, demonstrating its practical adaptability.

From a computational perspective, our proposed "Perception-Decision" architecture shares structural similarities with Ensemble Learning frameworks, specifically Stacking or Mixture of Experts (MoE). In such frameworks, the perception module can be viewed as a bank of base learners, while the decision-making module functions as a meta-learner. However, a critical distinction lies in the optimization objective of the base learners. Traditional ensemble methods typically employ weak classifiers trained at MinCE on the entire dataset. In contrast, our detectors are designed as "specialized experts" optimized for MaxDP. They do not aim to classify all samples correctly but rather to trigger only when a specific, high-confidence artifact is present. Furthermore, unlike the soft voting or weighted averaging often used in traditional ensembles, our decision module employs explicit logical reasoning via decision trees. This "hard" attribution of faults allows the system to identify fundamental forgery flaws rather than overfitting to complex decision boundaries, thereby leading to superior generalization on unseen attacks.

While this work establishes the foundational viability of our brain-inspired approach, we acknowledge several limitations and avenues for future research. First, the system's detection scope is fundamentally determined by the completeness of the chosen set of forgery clue detectors. While the system's detection scope relies on the current set of detectors, our architecture offers a dynamic defense mechanism against 'unknown unknowns.' Unlike conventional black-box models that may overfit to training data distributions, our approach focuses on fundamental physical anomalies, ensuring better generalization. Furthermore, facing the inevitable emergence of new forgery clues, our system's modularity serves as a critical advantage. It supports rapid iterative evolution: new specialized detectors can be developed and seamlessly integrated to cover emerging threats, ensuring the system continuously expands its coverage boundary without the need for catastrophic retraining. Future work could explore to optimize the combination of forgery clues that provides maximum coverage. The core objective is to identify a minimal set of clues capable of addressing the widest range of fake speech, thereby avoiding redundancy. More efficient clue detectors are also needed to provide more robust performance with expert design. Second, the static thresholds for each detector, while effective, were set based on the development set. Future work could explore dynamic or adaptive thresholding mechanisms to further optimize precision and reduce potential false positives in real-world scenarios. It is also worth noting that our framework establishes a solid foundation for finer-grained visual explainability. While the current system focuses on logic-level transparency (identifying which specific detector is activated), it can be readily extended with visualization techniques in future developments. Visualization methods[32] such as Class Activation Mapping (CAM)[33] can be applied to the specific activated detector. This would allow the system to not only report the presence of a frequency-domain artifact but also to visually highlight the precise region on the spectrogram that triggered the decision, offering a comprehensive diagnosis from high-level logic to low-level features.

In conclusion, our perception-decision machine offers a promising path toward building more robust, explainable, and adaptable speech security systems. Last but not the least, the multi-clue detection paradigm presented here has implications that extend far beyond speech, potentially benefiting any field facing diverse and evolving data. It is particularly well-suited for complex anomaly detection tasks where indicators of failure or attack are varied and subtle.

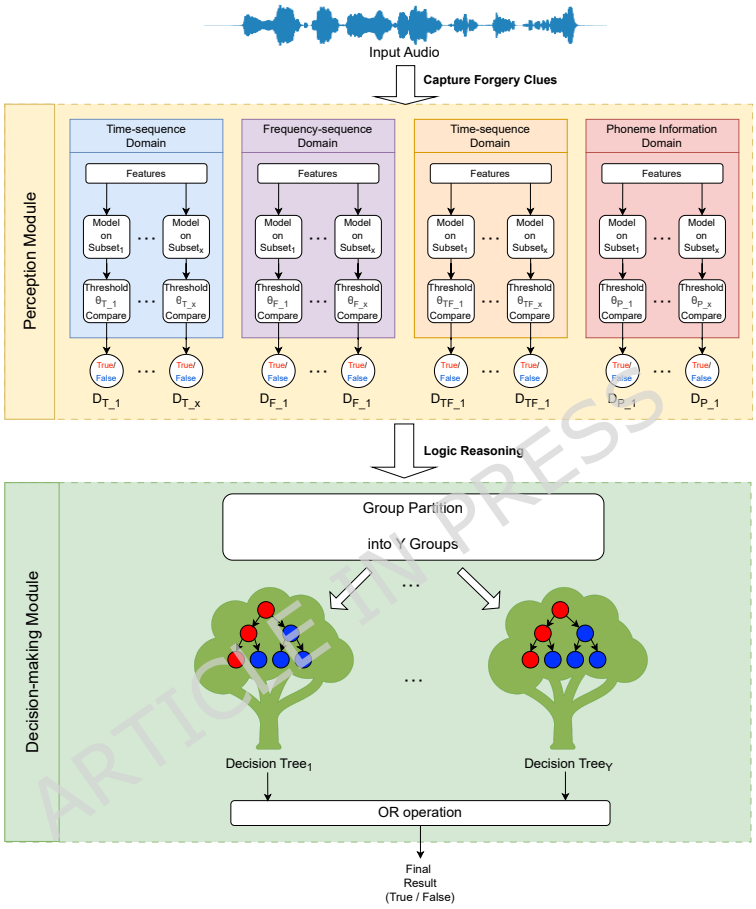
Methods

In this section, we provide the implementation detail of our method.

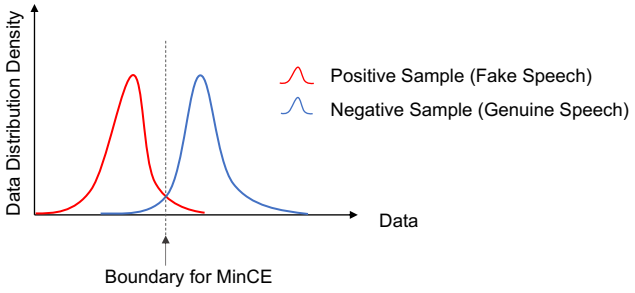
Framework of Perception-Decision Machine

The perception-decision machine consists of two modules: perception module and decision-making module, as shown in Figure 5a. The two modules are trained separately in order. The perception module processes speech audio using multiple detectors with maximized precision to identify a variety of forgery clues of fake speech. The output of a detector can be interpreted as

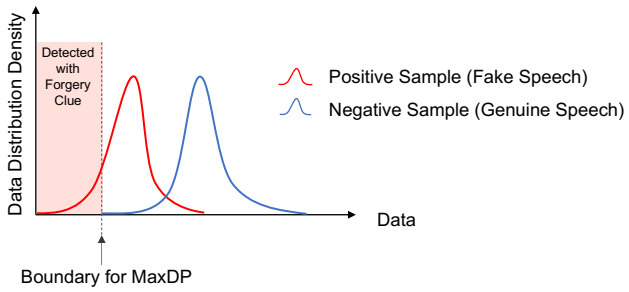
a presence or absence of a specific forgery clue. Then, the decision-making module relies on the outputs of detectors from the perception module, generating the final result through logical reasoning with forgery clue information. In practice, the detector's precision in identifying forgery clues tends to approach 100% in probability, but the detector is not guaranteed to be fully correct. Some genuine speech can also be falsely flagged with forgery clue. To further enhance overall detection performance, multiple perceived clues are considered in the decision-making process, which is organized with multiple decision trees.



(a) Perception-Decision Machine consists of perception module and decision-making module.



(b) Minimum Classification Error (MinCE).



(c) Maximum Detection Precision (MaxDP).

Figure 5. Perception-Decision Machine.

Perception Module

In the perception module, multiple detectors operate independently, with each one capturing a forgery clue and handling a particular detection sub-task. The detector takes the audio information as input and produces the output of Boolean value, indicating whether a specific forgery clue is present or not with maximum precision in probability. Different from previous classifiers aimed at Minimum Classification Error (MinCE as in Figure 5b) that refines decision boundaries to minimize both false positives (classifying genuine speech as fake) and false negatives (classifying fake speech as genuine), detectors are aimed at Maximum Detection Precision (MaxDP as in Figure 5c) that focuses on the precision of true positives, regardless of the missed detection in a single detector. Specifically, MaxDP is a thresholding strategy designed to minimize False Positives. Unlike MinCE which seeks a balanced decision boundary, MaxDP sets the threshold θ_{opt} at the extreme tail of the score distribution. The objective is to ensure that any detector activation (output=True) guarantees the presence of a forgery clue with nearly 100% confidence, essentially functioning as a high-precision filter.

Each detector comprises a computational model that calculates data pattern and a score threshold that is set to generate the detection result with maximum precision in probability. Unlike models of previous classifiers to compute two posterior probabilities of both fake and genuine speech classes, the computational model of detectors is designed like a score prediction in the regression tasks. The model calculates out a real number between 0 and 1, which represents the probability that the specific forgery clue exists within the input speech data. This regression task is structured such that the probability of the clue's presence in genuine speech is set to 0, indicating that the clue is absent. On the other hand, in fake speech that contains the specific clue, the probability is set to 1, indicating the absolute presence of the clue. If the model score exceeds the score threshold, the output value is of True; otherwise, it is of False. Specifically, the threshold is determined based on the detection precision for fake speech label on the development set, where we aim for 100% precision during the optimization phase. The score threshold is obtained through threshold search on the model scores from the training or development set, where the threshold is selected as the minimum score value that achieves a desired precision level. The threshold search is conducted on the model scores from the training or development set. where the threshold is selected as the minimum score value that achieves a desired precision level θ_{opt} represent the optimal threshold for a given detector. The threshold search process can be formalized as follows:

$$\theta_{opt} = \min\{\theta \in Score | Precision(Score) \geq p\} \quad subject\ to \quad Score \in [Score_{min}, Score_{max}], \tag{1}$$

where

- $Score$ is the set of score values that computed on the development set,
- $Precision(Score)$ is the detection precision of the detector when the Score value is applied as the detector threshold on the development set,
- p is the aimed detection precision which is set to the value of 1.0,
- $Score_{min}$ and $Score_{max}$ represent the minimum and maximum possible threshold score values on the development set.

Once the optimal threshold θ_{opt} is determined for each detector, it is used to make the outputs of the detector during the detection stage. The optimization of these thresholds ensures that each detector operates with maximum precision for its specific detection task, thus enhancing the overall performance of the multi-detector framework.

The detector models are flexible, allowing for different input features and computational structures. They function autonomously during the detection process and allow for parallel processing, which can further improve the efficiency and scalability of the system. The model design is based on the specific forgery clue with expert knowledge. Given that different fake speech generation methods tend to embed distinct types of forgery clues, we adopt a multi-path learning strategy, in which separate detectors are individually trained on fake speech generated by different algorithms. To construct a comprehensive set of potential forgery clues, we perform calculations from four different aspects: time-sequence domain, frequency domain, time-frequency domain, and phoneme information.

Perception in time-sequence domain

In fake speech, forgery clues are often revealed through inconsistencies in the smoothness and distribution of the signal waveform's sampling points. Different from genuine speech audios that reflect the natural fluctuations of the human voice and contain a fluid waveform with harmonious and smooth sampling points, fake speech may display abrupt transitions in different scales. These irregularities stem from the inherent limitations of fake speech generation methods. As a result, the audio may sound mechanical or artificial, with noticeable "jumps" in pitch, volume and other unharmonious noise.

To capture this kind of clues, the computational model of detectors applies RawNet2[12] model. It directly operates on raw audio waveforms, and utilizing SincNet[34] with sinc functions to parametrize a bank of band-pass filters in its first layer. These filter parameters are then learned automatically, which allows to capture waveform sampling point variations from dynamic scales, more effectively than traditional methods that rely on fixed filters. The detail structure is in Table 2a.

Block Type	Parameter Settings
Sinc filters	Conv1d(3,1,20) Maxpooling(3) BN & LeakyReLU
ResBlock × 2	BN & LeakyReLU Conv2d(3,1,20) BN & LeakyReLU Conv2d(3,1,20) Maxpooling(3) Feature Map Scaling
ResBlock × 3	BN & LeakyReLU Conv2d(3,1,128) BN & LeakyReLU Conv2d(3,1,128) Maxpooling(3) Feature Map Scaling
GRU	GRU(1024)
FC	Linear(1024,256) SeLU Linear(256,1) Sigmoid

(a) For forgery clues from time-sequence domain.

Block Type	Parameter Settings
Pre-trained Resnet-18	Conv(7,2,64) Maxpooling(2) ResBlock Conv2d(3,1,64) × 4 ResBlock Conv2d(3,1,128) × 4 ResBlock Conv2d(3,2,256) + Conv2d(3,1,256) × 3 ResBlock Conv2d(3,2,512) + Conv2d(3,1,512) × 3 Avgpooling(7) Linear(512,512)
FC	Linear(512,256) ReLU Linear(256,1) Sigmoid

(b) For forgery clues from frequency domain.

Block Type	Parameter Settings
Conv	Conv(3,1,32)
ResBlock	Conv2d(3,2,32) BN & LeakyReLU Conv2d(3,1,20) BN
FC	Linear(640,1024)
GRU × 2	GRU(1024)
Self-attention	Linear(1024,256) Tanh
FC	Linear(2048,256) Sigmoid Linear(256,1) Sigmoid

(c) For forgery clues from time-frequency domain.

Block Type	Parameter Settings
TDNN	Conv1d(5,1,1024) ReLU & BN
FTDNN × 7	Conv1d(2,1,256) × 2 Conv1d(1,1,1024) BN & ReLU
FTDNN	Conv1d(2,1,1024) × 3 BN & ReLU
FC	Linear(1024,2048) BN & Relu Concate(Mean + Std) Linear(4096,512) BN & ReLU Linear(512,1) Sigmoid

(d) For forgery clues from phoneme domain.

Table 2. The detail structure and settings of computational models for forgery clues from (a) time-sequence domain, (b) frequency domain, (c) time-frequency domain and (d) phoneme domain. Conv1d and Conv2d are convolutional layer of 1-dimension and 2-dimension, with (kernel number, kernel size, stride). BN is batch normalization operation.

Perception in frequency domain

The frequency domain represents signals in terms of their constituent frequencies base on short-time stationary analysis, rather than their time-based behavior, allowing for the analysis of how different frequency components contribute to the overall signal. Spectrogram is a visualized form of frequency domain and forgery clues of fake speech can be manifested in the spectrogram. Common signs of forgery include unnatural harmonic structures, where harmonics are too regular or evenly spaced, unlike the subtle variations in natural speech. Fake speech may also exhibit overly smooth spectral patterns, with formants that lack the natural fluctuations in human voices.

To capture this kind of clues, the detector input is the spectrogram of speech audio and the computational model applies ResNet-18[35,36] followed by two linear layers, a deep convolutional neural network (CNN) architecture. It includes convolutional

layers, pooling layers, and fully connected layers. It starts with an initial convolutional layer that processes the input image (in this case, the spectrogram of speech), followed by a series of residual blocks. Each residual block contains two 3x3 convolutional layers with batch normalization and ReLU activation functions. These blocks are designed to allow the network to learn residual mappings, which helps preserve information and enables more efficient training of deeper layers. The network concludes with a global average pooling layer, which reduces the spatial dimensions of the feature maps before passing them into a fully connected layer for classification. The residual connections help the model effectively learn hierarchical features—starting from basic spectral patterns like formants and harmonics, to more complex, high-level representations like phoneme transitions and timing anomalies. The detail structure is in Table 2b.

Perception in time-frequency domain

Forgery clues of fake speech can be exist in the joint observation from both time and frequency domain. They include misalignments or inconsistencies between the spectral and temporal components. In genuine speech, the frequency components evolve in harmony with the physiological process of pronunciation. However, in fake speech, some frequency components may be out of sync with the time, either appearing at incorrect moments or failing to appear when they should.

The 2D Discrete Cosine Transform (2D-DCT) is a transformation technique that combines both time and frequency information. First, a Discrete Cosine Transform (DCT) is applied in the time domain to capture the temporal characteristics of the speech signal. Then, a second DCT is performed in the frequency domain to reduce frequency redundancy within each column and to extract features associated with the formant group. The detector takes 2D-DCT as input features and applies a CNN-based model as computational model. The model contains one Residual Block, two bidirectional Gated Recurrent Units (GRUs) and a self-attentive pooling layer. The detail structure is in Table 2c.

Perception in phoneme information

The phonemes of speech are the smallest units that we can understand with speech content. Although fake speech generation algorithms are primarily designed to produce intelligible and clear speech content, forgery clues can still exist in the phoneme pronunciation level. The pronunciation of genuine speech is a result of the complex and dynamic coordination between various physiological components. And in fake speech, particularly in phonemes that require precise oral adjustments or intricate changes in airflow, the phonemes often lack the subtle, continuous modulation that characterizes natural pronunciation. This can manifest in several ways, such as overly precise or static articulation of certain sounds. For example, consonants that require complex tongue movements, like "t," "d," or "s," may sound sharp, clipped, or even overly harsh.

To extract features of phoneme information for detectors, we leverage XLSR (Cross-Lingual Speech Representation)[37], a pre-trained model designed for robust speech recognition across multiple languages. By learning on large-scale and multilingual speech data, XLSR captures nuanced phonetic features, enabling it to perform well even with diverse linguistic inputs. The detector takes the last embedding of XLSR as input features and the computational model is a CNN-based model with multi-head attention layer. The detail structure is in Table 2d.

Decision-making Module

The decision-making module is responsible for aggregating the outputs from the perception module and making the final judgment about whether the speech audio is fake or genuine. This module takes the binarized outputs of the individual clue detectors as inputs and conducts logical reasoning through decision trees and OR logic operation.

During this logical reasoning, the clue inputs to the decision-making module are first partitioned into multiple groups, each group constructing a decision tree. The group partition is learned from reinforcement learning. We apply Q-table in the learning. The state is a case of partition, represented by a vector whose dimension equals the number of detectors. Each element of the vector means a specific clue input is partitioned into a group. The action is to assign a clue input to another group, represented by a tuple. Its first element is the index of the clue input, which consists with the state vector index. And the second element is the group number which the clue input is assigned to. The action selecting is conducted through ε-greedy strategy. It contains exploration that chooses an action at random with probability ε and exploitation that selects an action with maximum Q-value in the current state with probability $1-\varepsilon$. The initial ε is set to 1 and will decrease with a decay ratio of 0.99. The reward of an action is based on information gain, the entropy difference between the old state and the new state. The end condition in each time of training is the The Q-value is stored in Q-table. The final group partition is the state that gets the maximum Q-value.

Each group of clue inputs undergoes an individual reasoning process using a decision tree model. The decision tree is learned with Gini Index metrics to identify the most relevant patterns in the clue information within each group. Specifically, the decision trees are implemented using the CART algorithm with Gini impurity as the split criterion and the depth determined automatically until leaves are pure. As the decision tree plays the role of logic reasoning, it enhances the explainability of the decision-making process.

The outputs of these individual decision trees are then combined with OR logic operation to make the final judgment. the OR logic operation conducts on variable-length elements, enabling incremental learning. Thus, the system can learn from new

fake speech data and incorporate new forgery clues over iterations. As new data is processed, the decision-making module adapts, continuously improving its ability to detect fake speech.

Experimental settings

The implementation of our method is carried out in Python. During detector training, we utilize the Adam optimizer with an initial learning rate of 10^{-4}, applying cosine annealing for learning rate scheduling. The batch size was configured at 200, and the model was trained over 10 epochs. Decision tree models were implemented using the SKLearn library, with Gini impurity chosen as the criterion for node splitting.

References

1. Kaur, N. & Singh, P. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artif. Intell. Rev.* **56**, 5837–5880 (2023).
2. Walczyna, T. & Piotrowski, Z. Overview of voice conversion methods based on deep learning. *Appl. sciences* **13**, 3100 (2023).
3. Shah, A. J. & Patil, H. A. Significance of lower frequency regions for audio deepfake detection. In *Proceedings of 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1–6 (IEEE, Macau, 2024).
4. Kirchhübel, C. & Brown, G. Spoofed speech from the perspective of a forensic phonetician. In *Proc. of 2022 Interspeech*, 1308–1312 (Incheon, 2022).
5. Sun, C., Jia, S., Hou, S. & Lyu, S. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 904–912 (IEEE, Vancouver, 2023).
6. Yang, J., Das, R. K. & Li, H. Significance of subband features for synthetic speech detection. *IEEE Transactions on Inf. Forensics Secur.* **15**, 2160–2170 (2020).
7. Zhang, Y., Wang, W. & Zhang, P. The effect of silence and dual-band fusion in anti-spoofing system. In *Proceedings of 2021 Interspeech*, 4279–4283 (International Speech Communication Association, Brno, 2021).
8. Todisco, M., Delgado, H. & Evans, N. W. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Proceedings of the 9th Speaker and Language Recognition Workshop Odyssey*, vol. 2016, 283–290, DOI: 10.21437/Odyssey.2016-4 (Bilbao, 2016).
9. Wu, Z., Das, R. K., Yang, J. & Li, H. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In *Proceedings of 2020 Interspeech*, 1101–1105 (International Speech Communication Association, Shanghai, 2020).
10. Wang, C. *et al.* Detection of cross-dataset fake audio based on prosodic and pronunciation features. In *Proceedings of 2023 Interspeech*, 3844–3848 (International Speech Communication Association, Dublin, 2023).
11. Tak, H., weon Jung, J., Patino, J., Todisco, M. & Evans, N. Graph Attention Networks for Anti-Spoofing. In *Proceedings of the 22nd Interspeech Conference*, 2356–2360, DOI: 10.21437/Interspeech.2021-993 (ISCA, Brno, 2021).
12. Tak, H. *et al.* End-to-end anti-spoofing with rawnet2. In *Proceedings of the 46th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6369–6373, DOI: 10.1109/ICASSP39728.2021.9414234 (IEEE, Toronto, 2021).
13. Chen, Y. *et al.* Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. In *Proc. of 2024 Interspeech*, 2720–2724 (International Speech Communication Association, Kos, 2024).
14. Chettri, B. *et al.* Ensemble models for spoofing detection in automatic speaker verification. In *Proceedings of the 20th Interspeech Conference*, 1018–1022, DOI: 10.21437/Interspeech.2019-2505 (ISCA, Graz, 2019).
15. Lavrentyeva, G. *et al.* STC Antispoofing Systems for the ASVspoof2019 Challenge. In *Proceedings of the 20th Interspeech Conference*, 1033–1037, DOI: 10.21437/Interspeech.2019-1768 (ISCA, Graz, 2019).
16. Tak, H., Patino, J., Nautsch, A., Evans, N. & Todisco, M. Spoofing Attack Detection Using the Non-Linear Fusion of Sub-Band Classifiers. In *Proceedings of the 21st Interspeech Conference*, 1106–1110, DOI: 10.21437/Interspeech.2020-1844 (ISCA, Shanghai, 2020).
17. Li, M., Ahmadiadli, Y. & Zhang, X.-P. A survey on speech deepfake detection. *ACM Comput. Surv.* **57**, 1–38 (2025).

18. Dhamyal, H., Ali, A., Qazi, I. A. & Raza, A. A. Using self attention dnns to discover phonemic features for audio deep fake detection. In *Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1178–1184 (IEEE, Cartagena, 2021).

19. Li, K., Lu, X., Akagi, M. & Unoki, M. Contributions of jitter and shimmer in the voice for fake audio detection. *IEEE Access* **11**, 84689–84698 (2023).

20. Parise, C. V. & Ernst, M. O. Correlation detection as a general mechanism for multisensory integration. *Nat. communications* **7**, 11543 (2016).

21. Pesnot Lerousseau, J., Parise, C. V., Ernst, M. O. & van Wassenhove, V. Multisensory correlation computations in the human brain identified by a time-resolved encoding model. *Nat. communications* **13**, 2489 (2022).

22. Rohlf, S., Li, L., Bruns, P. & Röder, B. Multisensory integration develops prior to crossmodal recalibration. *Curr. Biol.* **30**, 1726–1732 (2020).

23. Wang, X. *et al.* Asvspoof 2019: A large-scale public database of synthesized, conted and replayed speech. *Comput. Speech & Lang.* **64**, 101114, DOI: 10.1016/j.csl.2020.101114 (2020).

24. Liu, X. *et al.* Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, Lang. Process.* **31**, 2507–2522 (2023).

25. Ma, H. *et al.* Cfad: A chinese dataset for fake audio detection. *Speech Commun.* **164**, 103122 (2024).

26. Jung, J.-w. *et al.* Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proceedings of the 47th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6367–6371, DOI: 10.1109/ICASSP43922.2022.9747766 (IEEE, Singapore, 2022).

27. Tak, H. *et al.* Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *Proceedings of the 12th Speaker and Language Recognition Workshop Odyssey*, 112–119 (Beijing, 2022).

28. Liu, X. *et al.* Leveraging positional-related local-global dependency for synthetic speech detection. In *Proc. of 2023 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1–5 (IEEE, Rhodes Island, 2023).

29. Wang, S. *et al.* Memristor-based adaptive neuromorphic perception in unstructured environments. *Nat. Commun.* **15**, 4671 (2024).

30. Yu, F. *et al.* Brain-inspired multimodal hybrid neural network for robot place recognition. *Sci. Robotics* **8**, eabm6996 (2023).

31. Lin, X. *et al.* A brain-inspired computational model for spatio-temporal information processing. *Neural Networks* **143**, 74–87 (2021).

32. Grinberg, P., Kumar, A., Koppisetti, S. & Bharaj, G. What does an audio deepfake detector focus on? a study in the time domain. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2025).

33. Jung, H. & Oh, Y. Towards better explanations of class activation mapping. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1336–1344 (2021).

34. Ravanelli, M. & Bengio, Y. Speaker recognition from raw waveform with sincnet. In *Proceedings of 2018 IEEE Spoken Language Technology Workshop*, 1021–1028 (IEEE, Athens, 2018).

35. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (IEEE, Las Vegas, 2016).

36. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE conference on computer vision and pattern recognition*, 770–778 (IEEE, Las Vegas, 2016).

37. Babu, A. *et al.* Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of the 23rd Interspeech Conference*, 2278–2282, DOI: 10.21437/Interspeech.2022-143 (ISCA, Incheon, 2022).

Data availability

The fake speech data in this study is open source. ASVspoof2019 LA can be accessed at https://datashare.ed.ac.uk/handle/10283/3336. ASVspoof2021 LA can be accessed at https://zenodo.org/record/4837263. CFAD can be accessed at https://zenodo.org/records/8122764.

Acknowledgements

Acknowledged to Tianshan Talents Cultivation Program - Leading Talents for Scientific and Technological Innovation (No. 2024TSYCLJ0002).

Funding Declaration

This work was funded by Beijing Science and Technology Financial Innovation Support Project (Z221100001222005).

Author contributions statement

Chang Feng analyzed the data, performed the experiments and wrote the initial manuscript. Xiaolong Wu edited the figures, provided critical feedback and revised the manuscript. Hamdulla Askar and Mingxing Xu supervised the research. Lihong Cao reviewed and edited the final manuscript. Thomas Fang Zheng conceived the project, designed the study, reviewed and edited the final manuscript. All authors read and approved the final manuscript.

Appendix

Table Appendix1. Detailed thresholds and performance for each detector on the development set. MaxDP strategy ensures high precision across all detectors, regardless of the missed detection in a single detector.

Domain	Detector ID	Model Architecture	Threshold (θ_{opt})	Precision (%)	Recall (%)
Time-Sequence (SA)	SA_1	SincNet + ResBlock	0.4614	100.0	52.5
	SA_2	SincNet + ResBlock	0.1305	100.0	48.4
	SA_3	SincNet + ResBlock	0.2726	100.0	30.2
	SA_4	SincNet + ResBlock	0.4573	100.0	35.1
	SA_5	SincNet + ResBlock	0.2561	100.0	49.3
	SA_6	SincNet + ResBlock	0.9163	100.0	11.0
Frequency (FA)	FA_1	ResNet-18	0.9906	100.0	15.2
	FA_2	ResNet-18	0.8107	100.0	32.1
	FA_3	ResNet-18	0.9912	100.0	18.5
	FA_4	ResNet-18	0.9933	100.0	38.0
	FA_5	ResNet-18	0.8679	100.0	20.2
	FA_6	ResNet-18	0.9979	100.0	19.1
Time-Freq (PA)	PA_1	CNN + GRU	0.9944	100.0	12.5
	PA_2	CNN + GRU	0.9963	100.0	10.1
	PA_3	CNN + GRU	0.9966	100.0	15.3
	PA_4	CNN + GRU	0.9922	100.0	8.4
	PA_5	CNN + GRU	0.9995	100.0	23.0
	PA_6	CNN + GRU	0.9959	100.0	16.1
Phoneme (WA)	WA_1	XLSR + CNN	0.9993	100.0	30.5
	WA_2	XLSR + CNN	0.3345	100.0	48.1
	WA_3	XLSR + CNN	0.9985	100.0	32.4
	WA_4	XLSR + CNN	0.1319	100.0	55.6
	WA_5	XLSR + CNN	0.4596	100.0	29.8
	WA_6	XLSR + CNN	0.9998	100.0	15.0