VoiceWukong: Benchmarking Deepfake Voice Detection

Ziwei Yan[1]†, Yanjie Zhao[1]†, Haoyu Wang[1]‡

[1]*Huazhong University of Science and Technology*
†Ziwei Yan and Yanjie Zhao contributed equally to this work
‡Corresponding author: haoyuwang@hust.edu.cn

Abstract

With the rapid advancement of technologies like text-to-speech (TTS) and voice conversion (VC), detecting deepfake voices has become increasingly crucial. However, both academia and industry lack a comprehensive and intuitive benchmark for evaluating detectors. Existing datasets are limited in language diversity and lack many manipulations encountered in real-world production environments.

To fill this gap, we propose VoiceWukong, a benchmark designed to evaluate the performance of deepfake voice detectors. To build the dataset, we first collected deepfake voices generated by 19 advanced and widely recognized commercial tools and 15 open-source tools. We then created 38 data variants covering six types of manipulations, constructing the evaluation dataset for deepfake voice detection. VoiceWukong thus includes 265,200 English and 148,200 Chinese deepfake voice samples. Using VoiceWukong, we evaluated 12 state-of-the-art detectors. AASIST2 achieved the best equal error rate (EER) of 13.50%, while all others exceeded 20%. Our findings reveal that these detectors face significant challenges in real-world applications, with dramatically declining performance. In addition, we conducted a user study with more than 300 participants. The results are compared with the performance of the 12 detectors and a multimodel large language model (MLLM), i.e., Qwen2-Audio, where different detectors and humans exhibit varying identification capabilities for deepfake voices at different deception levels, while the MLLM demonstrates no detection ability at all. Furthermore, we provide a **leaderboard** for deepfake voice detection, publicly available at https://voicewukong.github.io.

1 Introduction

The rapid development of technologies such as TTS (Text-to-Speech) and VC (Voice Conversion) has brought great convenience to areas like entertainment and accessibility services. However, it's a double-edged sword. Illegal actors may exploit deepfake voices for various criminal activities. For example, in 2019, criminals used AI-based software to impersonate the voice of a U.K.-based energy firm's chief executive and requested a fraudulent transfer of €220,000 [3]. To counter the growing threats posed by deepfake voice technology, researchers are actively developing detection methods and creating open-source datasets for evaluation. For instance, traditional pipeline detection methods like LFCC-LCNN [64], the one-class learning-based detection method [74], emerging end-to-end detection models such as AASIST [29] and RawNet2 [54], as well as open-source deepfake voice datasets like ASVspoof [58, 68], have all garnered wide interest.

Unfortunately, academic deepfake voice detection methods often excel on specific datasets but fall short in real-world scenarios [43]. The rise of commercial tools and the latest generative models has produced increasingly convincing synthetic voices, outpacing current detection capabilities [45]. A key issue is the reliance on outdated or generic datasets for evaluation, which fail to reflect the sophistication of modern deepfake technologies. In practical applications, detectors struggle with poor generalization to unknown attacks and lack large-scale in-the-wild datasets [71]. Additionally, most methods focus solely on original content, overlooking the impact of post-processing manipulations (such as noise injection) on detection accuracy [67]. Given these challenges, there is a pressing need for a comprehensive benchmark to objectively evaluate various detection methods, thereby **bridging the gap between academic research and real-world applications**.

To address this gap, we introduce VoiceWukong, a comprehensive deepfake voice detection benchmark that incorporates various voice manipulations. VoiceWukong focuses on English and Chinese, the two most widely spoken languages globally, and features voices synthesized by advanced commercial tools and open-source models. We evaluated 12 state-of-the-art detection models, visually presenting their performance differences. Our fine-grained analysis of detector performance across different manipulations reveals potential avenues for optimization and improvement. Recognizing that **humans are the primary targets of deepfakes**, we conducted a user study involving over 300 participants. Based

on the results, we classified deepfake voices into three levels. We then analyzed the performance of the detectors at each level, comparing the detection capabilities of users versus automated systems for these synthesized voices.

Our main contributions can be summarized as follows:

- **A dataset addressing gaps.** Our dataset encompasses both English and Chinese languages, leveraging 19 advanced commercial tools and 15 open-source models. Through six types of manipulations, it has accumulated 38 data variants, resulting in a total of 265,200 English and 148,200 Chinese deepfake voice samples. To our knowledge, **this dataset is the first to extensively incorporate manipulation variants and compile the largest collection of commercially generated voice samples**.
- **A comprehensive benchmark.** We evaluated 12 advanced deepfake voice detectors using `VoiceWukong`. Results show that most detectors have an equal error rate (EER) above 20%, with three detectors exhibiting random performance on either the Chinese or English dataset. `AASIST2` [55] achieved the best EER (13.50% for English and 13.54% for Chinese), yet this falls significantly short of the 0.82% EER reported on its original evaluation dataset, underscoring the challenges these detectors face in real-world applications. We further compared detector performance and conducted a fine-grained analysis to identify specific manipulations that cause performance degradation for each detector.
- **A large-scale user study.** We conducted a user study involving over 300 participants to categorize deepfake voices into three levels of increasing difficulty (levels 0-2) based on their actual deception effectiveness. We then evaluated the performance of the 12 detectors and a multimodel large language model (MLLM), `Qwen2-audio` [13], across these levels. Results show that humans have false acceptance rates (FARs) of 18.97% for level 0 deepfakes in English and 4.20% in Chinese, outperforming most detectors. For level 2 deepfake voices, human FARs exceed 82% in both languages, falling behind most detectors. `Qwen2-audio` has an F1-Score of zero on the English dataset, indicating its inability to detect deepfake voices. We also examined human-focused features in deepfake voice detection to enhance detector-human collaboration in identifying synthetic audio.

2 Background and Related Work

2.1 Deepfake Voice Detection

Deepfake voice detectors primarily fall into two categories: traditional pipeline detectors and the increasingly researched end-to-end detectors [71]. The pipeline consists of a frontend feature extractor and a backend classifier. The classifier determines authenticity based on the features extracted by the feature extractor. Common features include spectral features represented by mel frequency cepstral coefficient (MFCC) [10], linear frequency cepstral coefficients (LFCC) [57], constant-Q transform (CQT) [12]; supervised embedding features [44]; and self-supervised embedding features represented by Wav2vec based features [55], XLS-R [7] based features [41]. Moreover, researchers have also attempted to explore some non-traditional features. Wang et al. [62] analyzed pop noise from close microphone speaking to detect deepfake voices. Doan et al. [18] detected deepfakes by evaluating the correlation between breathing, talking (speaking), and silence sounds. Common backend classifiers include traditional classifiers represented by GMM-based classifiers [15], and deep learning classifiers represented by ResNet [25] based classifiers [59], Res2Net [22] based classifiers [34], and DARTS [37] based classifiers [23].

End-to-end detectors have also received wide interest in the field. `RawNet2` [30] is a network designed for speech recognition and speaker verification that directly processes raw audio waveforms. Tak et al. [54] were the first to apply `RawNet2` to anti-spoofing. Wang et al. [65] proposed a joint optimization method based on the weighted additive angular margin loss to enhance the `RawNet2` based deepfake voice detector. Tak et al. [52] utilized the merit of graph attention networks (GATs) to learn the relationships between cues located in different sub-bands or different temporal intervals [56], proposing `RawGAT-ST`, which achieved excellent performance on ASVspoof2019. `RawGAT-ST` uses a pair of parallel graphs to simultaneously model temporal and spectral information, then merges them with element-wise multiplication. Jung et al. [29] proposed integrating these two heterogeneous graphs with heterogeneity-aware techniques and created the `AASIST`, which achieved superior performance on ASVspoof2019.

Unfortunately, most existing detectors are limited to pursuing performance on a single dataset, neglecting many challenges encountered in real-world applications. Ba et al. [6] highlighted these limitations in cross-language detection and proposed adaptation strategies. Zhang et al. [73] pointed out the insufficiency of models in adapting to unknown new attacks and conducted research on continual learning in deepfake voice detection. Wang et al. [63] and Wu et al. [67] considered the impact of manipulations on detector robustness, an aspect that most people have not taken into account.

2.2 Benchmarks and Datasets

Benchmarks. Recent research in deepfake detection has primarily focused on benchmarking face detection methods. Notable contributions include the CDDB benchmark by Li et al. [35], which simulates real-world scenarios, and the comprehensive evaluation by Deng et al. [16] using multiple generation methods and detection metrics. Pei et al. [46] provided a thorough survey and evaluation of deepfake face generation and detection techniques across various datasets

Table 1: The details of commonly used deepfake voice datasets. Note that, various datasets have issues such as single language and lack of manipulations.

Dataset	FoR	ASVspoof 2019		WaveFake	ASVspoof2021			ADD 2022			In-the-Wild
Year	2019	2019		2021	2021			2022			2022
Language	English	English		English & Japanese	English			Chinese			English
Corpus	A phrase dataset [2]	VCTK [60]		LJSpeech [1] & JSUT [50]	VCTK & Other[1]			AI-1,3,4 [9, 21, 49][2]			-
Subset	-	LA	PA	-	LA	PA	DF	LF	PF	FG-D	-
Types	TTS	TTS,VC	Replay	TTS	TTS,VC	Replay	TTS,VC	TTS,VC	PF[3]	TTS,VC	TTS
Goal	DD	ASV	ASV	DD	ASV	ASV	DD	DD	DD	DD	DD
Manipulation	-	-	-	-	Trans[4]	Noisy, Reverb	-	Noisy	-	-	Noisy
Commercial	6	0	-	0	0	-	0	-	-	-	-
Academic	1	17	-	7	13	-	>100	-	-	-	-

1: Other undisclosed corpora.
2: AI-1, AI-3, and AI-4 represent AISHELL-1, AISHELL-3, and AISHELL-4, respectively.
3: PF represents partially fake that generated by manipulating only a few words in the original bonafide utterances with real or synthesized voices.
4: The LA part of ASVspoof2021 transmitted the original voice through various telephone systems.

and sub-fields. In the voice domain, Zang et al. [72] introduced CtrSVDD, a large-scale benchmark for detecting singing voice synthesis models. **To the best of our knowledge, `VoiceWukong` is the first comprehensive and in-depth benchmark focusing on deepfake voice detection.**

Datasets. The commonly used evaluation datasets for deepfake voice detectors include FoR [48], ASVspoof2019 [58], WaveFake [20], ASVspoof2021 [68], ADD2022 [69], and In-the-Wild [43], as shown in Table 1. ASVspoof2019 is constructed for *automatic speaker verification (ASV)* tasks and includes a replay subset (ASVspoof2019-PA). It has received a lot of attention in the field of *deepfake detection (DD)*. ASVspoof2021 builds upon ASVspoof2019 by adding a section specifically for deepfake detection (ASVspoof2021-DF). Only WaveFake is constructed across multiple languages, but it does not focus on the most widely used languages. In-the-wild has a limited scope, focusing only on deepfake voices of celebrities and politicians. FoR is derived from seven open-source and commercial methods. Only a few datasets include manipulations like noise (ASVspoof2021, ADD2022, In-the-wild) and replay attacks (ASVspoof2019, ASVspoof2021). Overall, our dataset offers the broadest coverage of commercial tools, encompasses the widest range of voice manipulation variants, and targets the most representative languages.

2.3 Threat Model

The threat model in this study focuses on the malicious use of deepfake voice technology such as fraud and impersonation. Adversaries are assumed to have access to advanced commercial and open-source voice synthesis tools, enabling them to generate highly convincing synthetic voices in English and Chinese, and employ post-processing techniques to enhance realism and evade detection. The rapid advancement of voice synthesis technology presents a growing threat, often outpacing the development of detectors [45]. Current academic detectors may fail in the real world due to poor generalization and outdated datasets, creating a gap between lab results and practical effectiveness against sophisticated deepfake voices. These threats manifest in various scenarios, ranging from long-form impersonation attacks to brief voice commands (often consisting of a few words) targeting smart devices [26] and authentication systems. Our goal is to bridge this gap by providing a comprehensive benchmark that reflects real-world threats, evaluates state-of-the-art detection methods, and incorporates human perception in assessing deepfake voice detection effectiveness.

3 Benchmark Construction

In this section, we introduce the construction process of `VoiceWukong`, as illustrated in Figure 1. § 3.1 introduces the dataset construction process. § 3.2 presents our unified training and evaluation of detectors. § 3.3 details our large-scale user study. Finally, § 3.4 discusses the evaluation metrics.

3.1 Dataset Construction

3.1.1 Voice Collection

Generation Methods. Since TTS and VC are mainstream methods for generating deepfake voices [33], our dataset focuses on these two types. We collected 15 open-source generation models that are either prominent in the research field or have the highest star ratings on GitHub. Given that adversaries might use commercial tools to synthesize deepfake voices in real-world scenarios, we additionally collected 19 such tools capable of generating deepfake voices for research purposes and paid the necessary fees for their use. **We carefully examined the terms of service for each commercial tool to confirm their permissibility for research purposes.** `VoiceWukong` is a non-commercial resource, thereby safeguarding against any potential infringement of intellectual property rights. To our knowledge, our dataset involves the

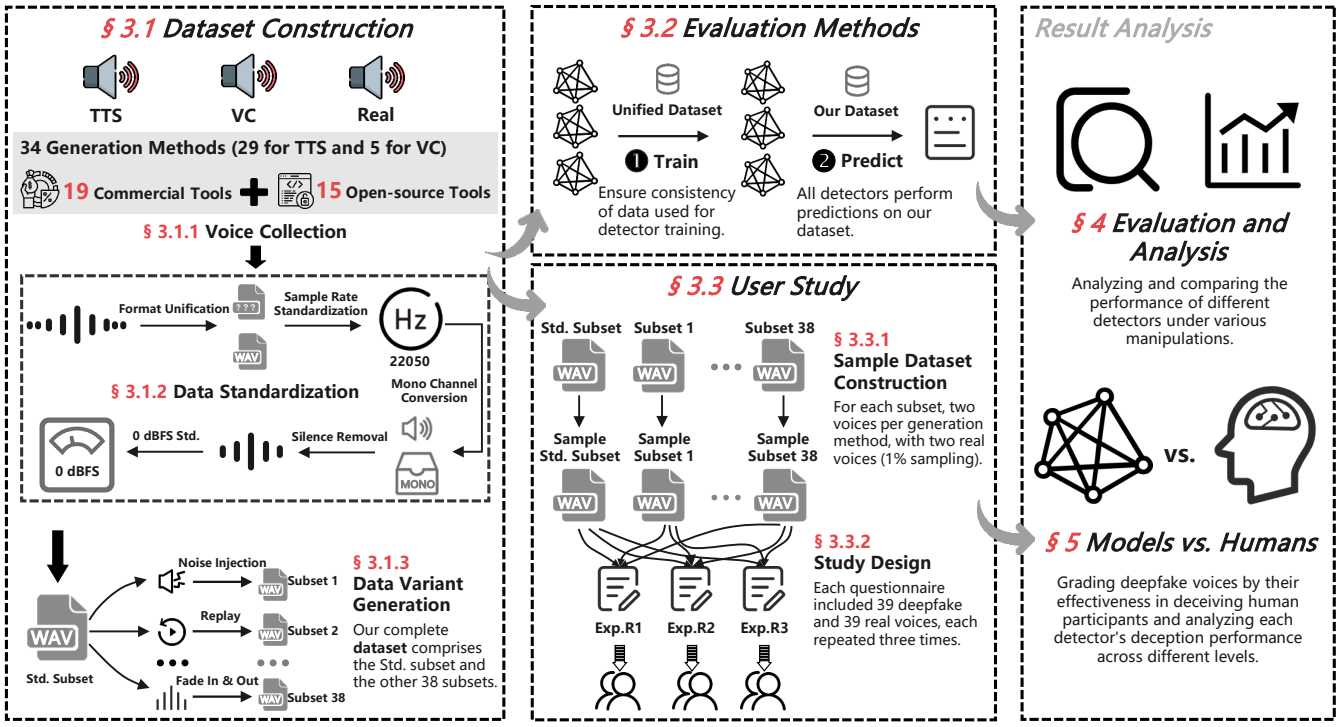


Figure 1: The overall workflow of the `VoiceWukong` benchmark construction.

most extensive range of commercial tools. The 34 methods (29 for TTS and 5 for VC) are detailed on our leaderboard [4], all supporting English and 19 supporting Chinese.

Generation Process. `VoiceWukong` encompasses English and Chinese, the two most widely used languages globally. Deepfake voice generation utilizes the English VCTK [60] and Chinese MAGICDATA [40] datasets. VCTK is widely used in voice cloning and conversion research, featuring diverse English textual content. MAGICDATA is a large-scale Mandarin speech dataset with extensive read speech data, covering domains like news, dialogues, and question-answering. After deduplication and length filtering (limiting to 5-30 words in English and 5-35 characters in Chinese accommodates brief voices as per the threat model in § 2.3), we selected 100 sentences each from VTCK and MAGICDATA for fixed-text generation, and extracted 3,400 English and 1,900 Chinese sentences for random-text generation. For each method, we produced 200 deepfake voices per supported language: 100 with fixed text and 100 with random text. Methods with 10 or more speakers yielded 10 fixed-text and 10 random-text voices per speaker from 10 speakers. For methods with fewer speakers, we allocated the 200 voices evenly among available speakers. Generation was automated for open-source models and API-based tools, while web-based tools required manual input. For the five open-source VC models, we provided corresponding original voices. In total, we collected 6,800 English deepfake voices across 34 methods and 3,800 Chinese deepfake voices across 19 Chinese-supporting methods.

Real Voice Collection. We also collected real voices from the VCTK and MAGICDATA datasets. To ensure data balance, we randomly selected text content not present in any deepfake voices and collected an equal number of real voices: 6,800 from VCTK and 3,800 from MAGICDATA.

3.1.2 Data Standardization

Inspired by previous research [48], the diversity of our dataset sources necessitates data standardization to eliminate biases. We implemented a standardized preprocessing pipeline for all voice samples, which included converting files to WAV format, resampling to 22,050 Hz, transforming to Mono channel, trimming silent segments from the beginning and end, and normalizing the volume to 0 dBFS.

Format Unification. Our collection methods include various open-source models and commercial tools, yielding non-uniform audio formats. To prevent format bias, we used `pydub`[1] to convert all files to WAV format, as it supports conversion between formats with minimal loss.

Sample Rate Standardization. Regarding sample rates, the voices from different sources also vary. To avoid sample rate effects on detectors, we standardized all data to a sample rate of 22,050 Hz. This rate, sufficient for the human speech spectrum (300 Hz to 3,400 Hz), offers adequate frequency

[1] https://github.com/jiaaro/pydub

resolution while saving space and improving processing efficiency compared to higher rates like 44,100 Hz and 48,000 Hz. We used the commonly applied `librosa`[2] for resampling all voice files to 22,050 Hz.

Mono Channel Conversion. Since the VCTK dataset uses two microphones for recording, we consistently used its data from `Microphone 2`. Additionally, the channel settings vary across different audio sources. To avoid the impact of different channels on the detectors, we used `librosa` to convert all voice files to Mono, eliminating bias.

Silence Removal. Real speech recordings often include silent segments at the beginning and end, which speech synthesis methods may not always replicate. To eliminate bias, we removed these silent parts from all voice files. A Python script was used to calculate the smoothed energy envelope of each voice, removing segments below the 20th percentile at the beginning and end.

Volume Normalization to 0 dBFS. When collecting voices from various commercial tools, some offer volume adjustment features while others do not. Moreover, the volume levels of voices generated by open-source models may vary. To eliminate the impact of volume differences on detection results, we normalize all voices to 0 dBFS.

3.1.3 Data Variant Generation

To cover various real-world scenarios that detectors might encounter, we applied additional manipulations to the standardized datasets. These manipulations reflect basic techniques malicious actors could use on deepfake voices and include six types: noise injection (NI), volume control (VC), time stretching (TS), sample rate changes (SR), replay (RE), and fade in & out (FD) effects. Except for replay, each type underwent fine-grained manipulations to create multiple variants.

Noise Injection [14, 48]. ESC-50 [47] is a widely recognized environmental sound classification dataset that divides sounds into five categories: animal sounds, natural soundscapes and water sounds, human (non-speech) sounds, interior/domestic sounds, and exterior/urban noises, each containing 400 five-second recordings. For each voice in our standardized dataset, we randomly selected a noise audio from each category and injected noise at Signal-to-Noise Ratios (SNRs) of 15 dB, 20 dB, and 25 dB. We also added Gaussian white noise to each voice using the same SNR settings.

Volume Control [66]. To examine if varying volume levels affect deepfake voice detection, we applied multiple volume adjustments to the standardized dataset. Since we were unaware of the adjustment outcomes, we conducted listening tests to ensure the voice content remained clearly audible. We set the lowest volume at 0.5 times and the highest at 1.5 times the standardized dataset, with intermediate levels at 0.75 and 1.25 times the original volume.

[2] https://github.com/librosa/librosa

Table 2: Overview of our dataset. Within the deepfake voices, random and fixed text accounts for half.

English Voices	530,400	**Fake Voices**	265,200	**Fixed-text Voices**	132,600
				Random-text Voices	132,600
		Real Voices	265,200	-	
Chinese Voices	296,400	**Fake Voices**	148,200	**Fixed-text Voices**	74,100
				Random-text Voices	74,100
		Real Voices	148,200	-	
Total	826,800	-			

Time Stretching [19]. Time stretching is a technique that changes the playback speed of audio without altering its other attributes, effectively adjusting the speech rate. We applied time stretching to the standardized dataset at 0.8, 0.9, 1.1, and 1.2 times the original speed to examine the impact of accelerated and decelerated playback on deepfake voice detection.

Voice Resampling. Generally, a higher sample rate [32] provides greater detail and quality in voice, making it sound more realistic and natural. However, higher sample rates also increase storage requirements. Therefore, generation tools often balance sample rates based on their needs. Lower sample rates lose some high-frequency information, potentially impacting detectors that rely on this data. We previously set 22,050 Hz as the sample rate for our standardized dataset and further resampled it at higher rates of 32,000 Hz and 44,100 Hz.

Replay [66]. Replay attacks [61] have long been exploited as a simple yet highly challenging method to test system robustness. To assess the robustness of different detectors against replay attacks, we re-recorded the standardized dataset. We played all voices at maximum volume on a Lenovo Xiaoxin Pro14 2023 laptop and recorded them simultaneously with a Deli 14870 omnidirectional microphone placed 1 meter away at the same height. The recordings were conducted in a quiet, unoccupied indoor environment.

Fade In & Out. Fade in & out are common editing techniques that create smooth transitions and more natural voice connections [38, 39]. Fade in gradually increases the volume from zero at the beginning of a voice clip, while fade out gradually decreases it to zero at the end.We applied the fades to each voice in the standardized dataset with linear, logarithmic, and exponential shapes at 0.1, 0.2, and 0.3 ratios.

Using six types of manipulations, **we generated 38 variants of the standardized (std.) subset, resulting in a total of 39 subsets**. As shown in Table 2, our dataset construction process yielded 265,200 English and 148,200 Chinese deepfake voices, along with an equal number of real voices.

3.2 Evaluation Methods

We now describe the evaluation of existing deepfake voice detectors using our constructed dataset. Our detector selection process is detailed in § 4.1. All detectors are trained on the ASVspoof2019-LA dataset (if necessary). We use published pre-trained models when available, or retrain models using the original authors' hyperparameters and settings, selecting

the best-performing checkpoint for evaluation. This approach preserves each detector's optimal performance as determined by its creators. During evaluation, we maintain the input sequence length specified by the original authors, rather than fixing the voice sampling duration, due to differences in sample rates between our dataset and ASVspoof2019. This ensures consistency between training and evaluation sequence lengths, enabling fair comparisons across detectors. This methodology allows us to compare each detector's performance while maintaining its peak capabilities and avoiding inconsistencies that could arise from fixed sampling durations.

3.3 User Study

To examine the real-world deception effectiveness of deepfake voices generated by the selected methods and manipulation variants, we conducted a user study with 318 participants. Based on the deception performance of these voices in the study, we divided the dataset into three levels to enable further evaluation of detectors in § 5.2. This section outlines the data sampling, study design, and grading method, while ethics and participant details are discussed in the Ethics Considerations section and Appendix A.

3.3.1 Sample Dataset Construction

Given our dataset's large scale, we sampled 1% for the user study, balancing practicality with comprehensive representation. We focused on the random text portion to avoid potential bias from repeated content in the fixed text samples. As illustrated in Figure 1, our sampling process was: 1) For each subset, we randomly selected two random-text deepfake voices per generation method, ensuring unique texts across subsets. 2) We then sampled an equal number of real voices from the same subset. This approach resulted in 38 Chinese deepfake and 38 real voices (corresponding to the 19 generation methods supporting Chinese), as well as 68 English deepfake and 68 real voices (34 generation methods) per subset. With a total of 39 subsets, our final **sample dataset** comprised 2,964 Chinese voices and 5,304 English voices.

3.3.2 Study Design

Questionnaire Setup. To maintain full control over the dataset, we set up a questionnaire on our server and provided each participant with a unique link to access the survey. At the start, users were required to read instructions that explicitly directed them to use headphones to minimize environmental noise and interference. Participants could only proceed after reading, understanding, and agreeing to these instructions. Furthermore, the response time for the entire questionnaire was strictly limited. Questionnaires completed in less than 25 minutes or more than two hours were considered invalid. For each subset within the **sample dataset**, we randomly paired one deepfake voice with one real voice. To construct each questionnaire, we randomly selected one unused pair from each subset, resulting in 78 voices per questionnaire. This ensured that every voice from the sample dataset appeared in each questionnaire round. After one round of questionnaire generation, we produced 38 different questionnaires for the Chinese dataset and 68 for the English dataset. We conducted three rounds of data collection, ultimately generating 114 Chinese questionnaires and 204 English questionnaires.

Tasks. The user study is structured around three distinct tasks. In the first task, participants are presented with randomly selected voice samples and asked to determine their origin. As illustrated in Figure 5 in the Appendix B, users listen to each voice and must categorize it as either "Human" or "Not human". They are required to make a selection before proceeding and are not permitted to revise their previous choices. Upon completion of the initial assessment, the second task commences. Participants revisit each deepfake voice sample, with their earlier judgments displayed as a reminder. They are then tasked with evaluating the generation quality on a three-point scale: `1` indicates an easily detectable deepfake, `2` suggests a deepfake identifiable upon close scrutiny, and `3` represents a voice indistinguishable from a genuine human recording. Following this evaluation, participants are prompted to identify the factors that influenced their decision-making process. They can select from predefined options such as volume and background noise or provide custom responses. The third task serves as an attention check for participants. Additionally, attention tests are embedded once within both the first and second tasks. These tests require participants to listen to a specific recording and answer content-related questions. Incorrect responses result in immediate termination of the questionnaire to prevent random guessing, and these participants are subsequently disqualified from the study.

Deepfake Voice Deception Grading. We developed a systematic approach to grade deepfake voices based on their effectiveness in deceiving human participants. The grading process for deepfake voices produced by a specific tool under particular manipulation conditions was as follows. Our study consisted of three experimental rounds (as illustrated in Figure 1), each considered a separate experiment within our overall investigation. For each round, we collected responses from 38 Chinese and 68 English questionnaires. As noted in § 3.3.1, each generation method was represented by two deepfake voice samples from each subset. We analyzed the deception outcomes for each experimental round, classifying them into three distinct scenarios **corresponding to different deception levels (0 to 2)**: no voices successfully deceiving humans (level 0), one voice successfully deceiving humans (level 1), or two voices successfully deceiving humans (level 2). The final grade for a generation method under specific conditions was determined by the majority outcome across the three experimental rounds. In cases of evenly distributed

results (i.e., one round each at levels 0, 1, and 2, regardless of order), we assigned a final level of 1.

3.4 Metrics

In VoiceWukong, we employ a comprehensive set of quantitative metrics to analyze the performance of various detectors. These metrics include False Acceptance Rate (FAR), False Rejection Rate (FRR), Equal Error Rate (EER) [42], Accuracy (ACC), F1-score, and Area Under the Curve (AUC). This diverse array of metrics provides a thorough evaluation of each detector's prediction results, offering insights into different aspects of their performance.

Let θ be the score threshold at which the detector classifies a voice as genuine. The $FRR(\theta)$ and the $FAR(\theta)$ of the detector at the threshold θ are defined as follows:

$$FAR(\theta) = \frac{\sum\{fake\ voices\ with\ score > \theta\}}{\sum\{fake\ voices\}} \quad (1)$$

$$FRR(\theta) = \frac{\sum\{real\ voices\ with\ score < \theta\}}{\sum\{genuine\ voices\}} \quad (2)$$

Equation 1 and Equation 2 are respectively decreasing and increasing functions of the threshold θ. The EER is defined in Equation 3 as the error rate at the specific threshold θ_{eer}, where $FAR(\theta_{eer})$ and $FRR(\theta_{eer})$ are equal.

$$EER = FAR(\theta_{eer}) = FRR(\theta_{eer}) \quad (3)$$

$$TAR(\theta) = \frac{\sum\{real\ voices\ with\ score > \theta\}}{\sum\{genuine\ voices\}} \quad (4)$$

The AUC represents the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve describes the trade-off between the False Acceptance Rate (FAR) and the True Acceptance Rate (TAR) (defined in Equation 4) of a detector at various decision thresholds (θ). AUC values range from 0 to 1, with higher values indicating better detector performance. An AUC close to 0.5 suggests the detector's performance is comparable to random guessing. ACC describes the detector's correct prediction rate on the dataset, as defined in Equation 5. The F1-Score (defined in Equation 6) reflects the detector's sensitivity to FAR and FRR. Together with ACC, it provides a more comprehensive analysis of detector performance on manipulation subsets (§ 4).

$$ACC(\theta) = \frac{\sum\{voices\ with\ correct\ score\}}{\sum\{\text{fake voices}\} + \sum\{\text{real voices}\}} \quad (5)$$

$$F1\text{-}Score(\theta) = \frac{(1 - FAR(\theta)) + (1 - FRR(\theta))}{2 \cdot (1 - FAR(\theta)) \cdot (1 - FRR(\theta))} \quad (6)$$

In the subsequent sections of this paper, all discussed metrics are under the condition of $\theta = \theta_{eer}$.

4 Evaluation and Analysis

4.1 Evaluated Detectors

We collected 12 open-source detectors that have gained significant attention in deepfake voice detection or demonstrated excellent performance in their original publications. Our focus was primarily on end-to-end detectors rather than traditional pipeline detectors, as the latter often require complex and time-intensive feature extraction processes for large-scale datasets. The selected detectors include AASIST [29] and RawNet2 [54], used as baselines in well-known challenges [31,68–70], along with RawBoost [53], OC-Softmax [74], RawGAT-ST [52], SAMO [17], Res-TSSDNet [28], RawNet2-Vocoder [51], AASIST2 [55], Raw PC-DARTS [24] and the latest detectors RawBMamba [11] and CLAD [67]. We evaluate all detectors on our dataset as described in § 3.2.

4.2 Evaluation Results

We examine the overall performance of various detectors. Table 3 shows the EER values for all detectors on our dataset. On the English dataset, EERs range from 13.50% to 50.01%, and on the Chinese dataset, from 13.54% to 51.88%. The best-performing detector is AASIST2, with the lowest EER on both datasets: 13.50% for English and 13.54% for Chinese, while all other detectors have EERs above 20%. The worst performer on the English dataset is Res-TSSDNet (50.01% EER), while on the Chinese dataset, AASIST has the highest EER at 51.88%. Most detectors perform better on the English dataset than on the Chinese dataset, likely due to their training on English data. EER differences between datasets indicate varying adaptability in cross-lingual detection. AASIST2 has the smallest EER difference (-0.04%) between the English and Chinese datasets, while AASIST shows the largest (-23.69%). SAMO also has a significant gap, with its English EER 23.15 percentage points lower than its Chinese EER, second only to AASIST. Interestingly, OC-Softmax performs better on the Chinese dataset, with its EER 3.51 percentage points lower. Res-TSSDNet has EER values close to 50% on both datasets, indicating poor performance across languages. On the English dataset, aside from AASIST2 (best) and Res-TSSDNet (worst), most detectors have similar EERs around 25%. However, on the Chinese dataset, performance differences are more pronounced, highlighting variations in cross-lingual detection capabilities among detectors.

We further assess each detector's robustness using AUC scores across different discrimination thresholds. Figure 2 presents the AUC scores and ROC curves for each detector's predictions. As shown in Figure 2a and Figure 2d, only AASIST2 achieved AUC scores above 0.9 on both language datasets. On the English dataset, four detectors (CLAD,

Table 3: The EER (%) values of all detectors, the difference between EER values on the English and Chinese datasets, and their FARs (%), FRRs (%) at the equal error point on the replay subset, as well as the difference between these two values. The numerical values appear in the following order: EER on the English dataset, EER on the Chinese dataset, the difference between these two, FAR on the English replay subdataset, FRR on the English replay subset, the difference between FAR and FRR on the English replay subset, FAR on the Chinese replay subset, FRR on the Chinese replay subset, and the difference between FAR and FRR on the Chinese replay subset.

	AASIST [29]	**RawNet2 [54]**	**RawBoost [53]**	**OC-Softmax [74]**	**RawGAT-ST [52]**	**CLAD [67]**
EN-EER	28.16	25.22	23.48	32.37	25.63	20.00
CN-EER	**51.88**	33.23	26.55	28.86	35.49	33.05
Difference	**-23.69**	-8.01	-3.07	**+3.51**	-9.87	-13.05
EN-FAR	17.88	80.74	47.24	99.47	15.25	44.24
EN-FRR	55.34	12.65	32.63	0.00	68.28	17.99
Difference	-37.45	+68.09	+14.60	+99.47	-53.03	+17.99
CN-FAR	54.97	93.95	53.68	93.97	30.21	77.61
CN-FRR	24.13	2.39	18.37	1.34	50.29	10.97
Difference	+30.84	+91.55	+35.32	+92.63	-20.08	+66.63

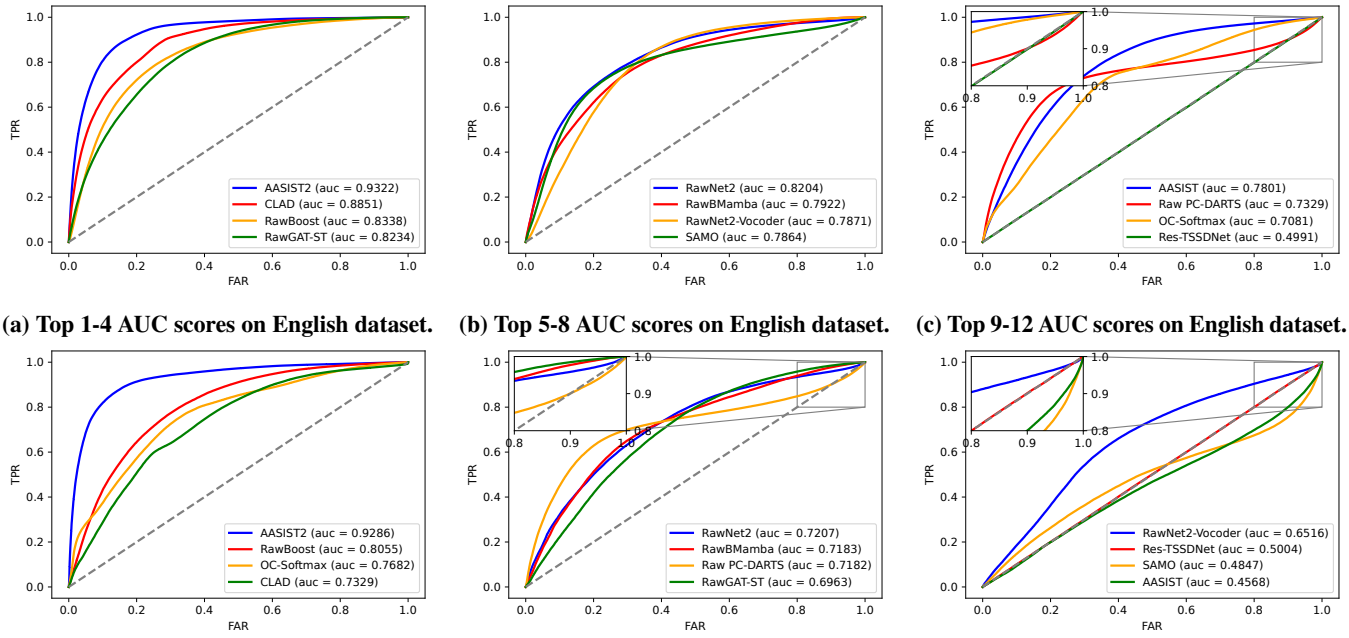
	Res-TSSDNet [28]	**RawNet2-Vocoder [51]**	**AASIST2 [55]**	**Raw PC-DARTS [24]**	**RawBMamba [11]**	**SAMO [11]**
EN-EER	**50.06**	27.54	**13.50**	27.93	27.43	25.57
CN-EER	49.91	37.01	**13.54**	29.99	32.48	48.72
Difference	+0.15	-9.47	-0.04	-2.07	-5.02	-23.15
EN-FAR	91.26	61.10	5.62	99.65	93.51	29.71
EN-FRR	9.03	32.74	72.59	0.00	1.37	26.25
Difference	+82.23	+28.37	-66.97	+99.64	+92.15	-14.72
CN-FAR	49.58	85.84	13.39	99.53	94.74	46.13
CN-FRR	49.97	9.74	33.79	0.05	4.84	37.13
Difference	-0.39	+76.10	-20.39	+99.47	+89.89	+9.00

RawBoost, RawGAT-ST, and RawNet2) have AUC scores between 0.8 and 0.9, while all other detectors, except Res-TSSDNet, fall between 0.7 and 0.8. On the Chinese dataset, only RawBoost has an AUC score above 0.8, while five other detectors (OC-Softmax, CLAD, RawNet2, RawBmamba and RawNet2-Vocoder) scoring between 0.7 and 0.8. Notably, AASIST2, CLAD, and RawBoost consistently rank in the top four for AUC scores in both languages. Res-TSSDNet performs poorly on both datasets, with AUC scores close to 0.5 (Figure 2c and Figure 2f). On the Chinese dataset, SAMO and AASIST also show similarly poor performance. The ROC curves for these detectors nearly overlap with the random prediction line, indicating extremely poor performance. Interestingly, Raw PC-DARTS exhibits reverse prediction behavior under certain thresholds on both datasets (Figure 2c and Figure 2e), as its ROC curves cross the random prediction line. Figure 2f shows that SAMO and AASIST also display this phenomenon on the Chinese dataset.

Among the 12 evaluated detectors, AASIST2 demonstrates the best overall performance and robustness. However, its EER on our dataset significantly underperforms compared to the 0.82% reported in the original paper. Other detectors also show notable discrepancies from their originally reported performances, for instance, RawBoost's best EER was 5.31% in the original paper but reached 23.48% in VoiceWukong. These findings raise concerns about the practical effectiveness of deepfake voice detection in real-world scenarios.

4.3 Effect of Manipulations

We analyze the detectors' performance across various manipulated subsets of our dataset, using each detector's performance on the standardized (std.) subset as a baseline. This approach helps identify specific differences under various manipulations and reveals potential optimization methods. The ACCs, F1-Scores, FARs and FRRs for all detectors at θ_{eer} on various subsets can be viewed on our leaderboard [4]. Due to space



(a) Top 1-4 AUC scores on English dataset. **(b) Top 5-8 AUC scores on English dataset.** **(c) Top 9-12 AUC scores on English dataset.**

(d) Top 1-4 AUC scores on Chinese dataset. **(e) Top 5-8 AUC scores on Chinese dataset.** **(f) Top 9-12 AUC scores on Chinese dataset.**

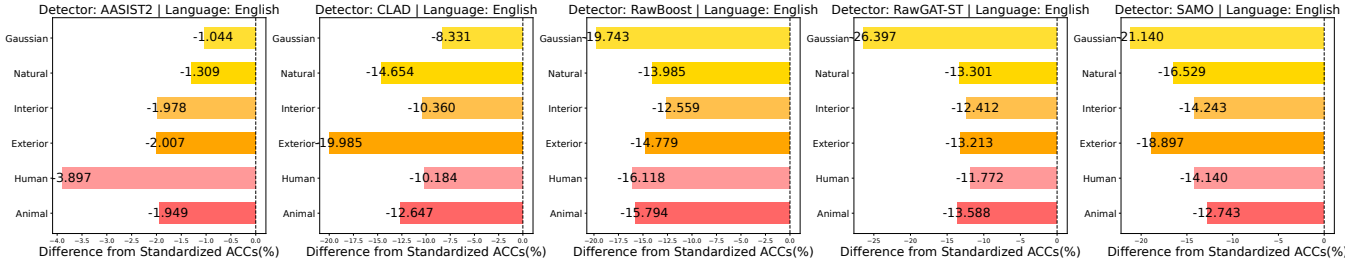
Figure 2: ROC curves for all detectors. (a), (b), and (c) show the performance of various detectors on the English dataset, while (d), (e), and (f) show their performance on the Chinese dataset.

constraints, we focus on the top four detectors by AUC score in both languages, while discussing the remaining detectors only for their unique performance characteristics.

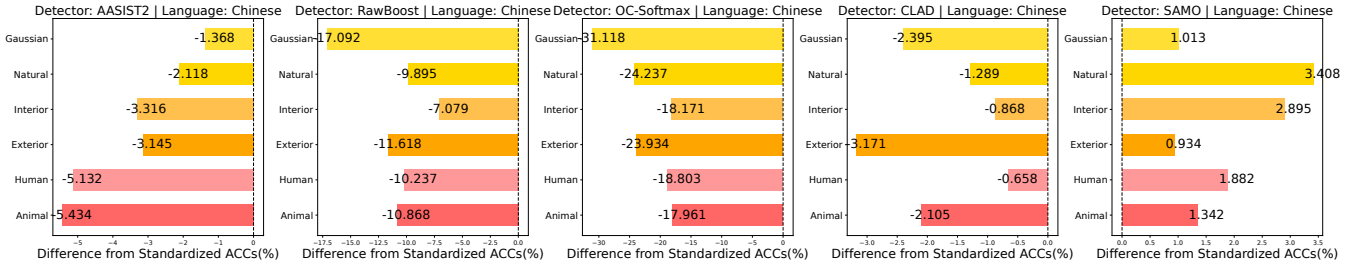
Analysis of the std. subset performance reveals AASIST2 as the standout performer, achieving ACCs above 90% on both language datasets. We focused on its FAR and FRR to evaluate its true potential. In English, it recorded a FAR of 13.97% and a FRR of 3.56%, while in Chinese, the rates were 11.68% and 8.18%, respectively. Despite leading in overall performance, AASIST2's FAR exceeding 10% in both languages indicates that there is still room for potential risk in real-world applications. On the English dataset, most detectors exceed 80% ACC, with CLAD peaking at 85.25%. Only RawNet2 and AASIST fall slightly below 80%. The Chinese dataset shows a general performance decline, except for OC-Softmax. AASIST experiences the most severe drop to 48.72% ACC. Notably, CLAD achieves only 68.67% ACC on the Chinese std. subset, despite performing well on the English std. subset. These results align with the AUC performance evaluation in § 4.2, highlighting varied cross-lingual detection capabilities.

Effect of Noise Injection. Figure 3 illustrates ACC differences between 15 dB noise-injected and std. subsets for the top four detectors. All detectors show ACC declines across both datasets, with some interesting exceptions. SAMO's ACC drops significantly on the English dataset but slightly increases on the Chinese dataset after noise injection. Specifically, in English, the FRR rises from 6.56% to over 50%, while the FAR decreases from 29.59% to below 10.56%. In Chinese, the FRR increases from 40.32% to over 62.58%, and the FAR drops from 56.74% to below 31.79%. This suggests that under this manipulation, SAMO is more likely to classify voices as deepfake. Different noise types impact detectors variably. For instance, Human noise notably affects AASIST, while Gaussian noise has the least impact. Conversely, RawBoost is most affected by Gaussian noise and least by Interior noise. The top four detectors show similar patterns of noise impact across languages. AASIST2 is least affected by noise on the English dataset, while CLAD is most resilient on the Chinese dataset. Overall, **when facing noise injection, most detectors may experience a very significant performance decline**. Only AASIST2 is the least affected, with a maximum ACC drop of 5.434% on the Chinese dataset with Animal-type noise injection (the FAR has only decreased by 3.08%). The impact varies across different noise types and detectors, underscoring the complexity of noise effects on deepfake voice detection.

The higher the SNR, the smaller the impact of noise on the voice, and the higher the quality of the voice. To explore the impact of different SNRs on the performance of the detectors, in the manually injected noise experiments, we set different SNRs of 15dB, 20dB, and 25dB for each type of noise. Figure 6 in the Appendix C shows the ACC of the top four AUC-ranked detectors under different SNRs. On the English dataset, for all the detectors evaluated in this work except Res-TSSDNet, the prediction ACC increases with the increase of SNR. On the Chinese dataset, except for AASIST,



(a) ACC differences for English dataset: the top four AUC-ranked detectors and SAMO on SNR 15dB noise-injected vs. std. subsets



(b) ACC differences for Chinese dataset: the top four AUC-ranked detectors and SAMO on SNR 15dB noise-injected vs. std. subsets

Figure 3: ACC differences: the top four AUC-ranked detectors and SAMO on SNR 15dB noise-injected vs. std. subsets

`Res-TSSDnet`, and `SAMO`, the ACCs of the remaining detectors also increase with the increase of the SNR, and only `RawBmamba` shows the opposite trend under Gaussian noise injection. The main reason for this trend is that with the injection of Gaussian noise, its FARs experienced a significant increase, reaching as high as 93.1% when the SNR is at 15dB. The impact of SNRs on different detectors varies. As shown in Figure 6a, the ACC change of `AASIST2` is minimal under different SNRs, while `CLAD`, `RawBoost`, and `RawGat-ST` show more significant changes. The difference in data language can also lead to different performance of the detectors under changes in SNRs. On the English dataset, `CLAD` shows a significant change in ACC, but on the Chinese dataset, this change is minimal.

Effect of Volume Control. For all detectors, the impact of volume control on detection performance shows no clear linear relationship based on the ACCs and F1-Scores results. Generally, volume control mainly adjusts the amplitude of the voice without significantly affecting other important features, such as spectral characteristics and timbre. Figure 7 in the Appendix C shows the cosine similarity of spectrograms calculated for the same voice at different volume levels, further confirming this point. Detectors typically rely on specific features to identify deepfake voices. During data collection, the range of volume adjustments was determined through human listening tests to ensure that the voice content remained audible. Therefore, the impact on features is even more limited, resulting in relatively stable performance across all detectors when facing volume control.

Effect of Time Stretching. Figure 4 shows the ACCs of various detectors under different factors of time stretching. Although the `AASIST2` performs exceptionally well in noise injection scenarios, it exhibits the most significant performance degradation among the top four AUC-ranked detectors when faced with time-stretching manipulations. Moreover, `AASIST2`'s detection performance declines with increasing time stretching, as the FAR decreases while the FRR increases. This effect is most pronounced in the English dataset, where the FRR rises to 79.51% and the FAR drops to 1.19% at a stretching factor of 1.1. In the English dataset, the performance of the other three detectors remains relatively stable, with `CLAD` being the most consistent detector. Similarly, in the Chinese dataset, the three detectors other than `AASIST2` also demonstrate stable performance, with the `OC-Softmax` showing the best results. `AASIST2` relies on the frontend features of wav2vec 2.0 [8], which is sensitive to temporal information. Time-stretching manipulations can cause shifts or distortions in some key temporally related features. Consequently, `AASIST2` has relatively weak resistance to time stretching interference, resulting in a noticeable decline in performance.

Effect of Voice Resampling. In terms of ACCs and F1-Scores, all detectors, except for `Raw PC-DARTS` and those detectors with near-random prediction AUC on their respective language datasets, experienced severe performance degradation after voice resampling. In the English dataset, `CLAD` and `RawNet2` saw their F1-Scores drop below 30% (29.75% and 25.85% respectively) at a sample rate of 44100 Hz. In contrast, `Raw PC-DARTS` maintained ACCs and F1-Scores close to the std. subset for both English and Chinese datasets. **When resampling from the std. subset to higher sampling rates, more feature information is lost in the same length of sampling sequence.** We speculate that this is the main reason for

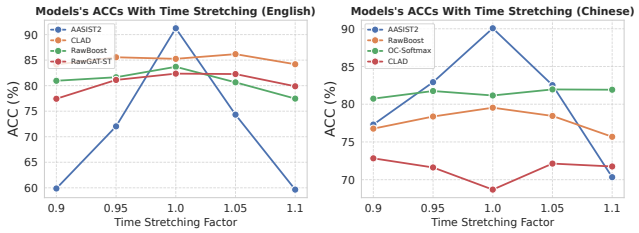


Figure 4: The ACC changes of the top four AUC-ranked detectors under different time stretching factors. AASIST2 showed a performance that was strikingly different from previous evaluations. Its performance fluctuated severely under time stretching.

the performance decline of most detectors when facing resampling. Raw PC-DARTS employs learnable Sinc filters, which suggests that it can still extract key features when confronted with different sample rates. This also explains its outstanding performance stability when dealing with resampled datasets.

Effect of Replay. As observed from the results, all detectors show a significant decrease in ACCs and F1-Scores when faced with replay subset. We further analyzed this using the FARs and FRRs of each detector. Table 3 presents the FARs, FRRs, and the difference between them for each detector on the replay subset. RawNet2, Raw PC-DARTS, OC-Softmax, and RawBMamba exhibit FAR values significantly higher than their FRR values in both languages. This indicates that these detectors are highly likely to classify replayed voices as real voices, demonstrating a clear lack of defense against easily implementable replay attacks. CLAD shows a FAR 17.99 percentage points higher than its FRR in the English dataset, but this difference increases to 66.63 in the Chinese dataset, suggesting the detector's vulnerability to replay attacks. In contrast, AASIST2 and RawGAT-ST have FAR values noticeably lower than their FRR values in both languages. This implies that these two detectors tend to classify replayed voices as deepfake voices, making them more suitable for hindering deepfake voices that have undergone replay operations.

Effect of Fade In & Out. Observing the statistical data in the fade-in and fade-out parts of the results, we observed that on the English dataset, AASIST2, CLAD, and RooBoost show no significant changes in ACCs when facing linear and exponential fade-in and fade-out. The proportion of fade influences the ACC, but the impact is minimal. However, when facing logarithmic fade, the performance decline of these three detectors is more noticeable. From the changes in FARs and FRRs, we can infer that the detectors tend to classify deepfake voices as real voices. RawGAT-ST consistently performs the worst among the top four AUC-ranked detectors. Its ACC declines noticeably in all three shapes of fade, with the decline becoming more pronounced as the proportion increases. It also shows a more severe tendency to misclassify deepfake voices as real voices. On the Chinese dataset, most detectors exhibited lower performance compared to their results on the English dataset, and the overall trend of changes is consistent with that of the English dataset. Interestingly, OC-Softmax shows surprisingly stable ACCs on the Chinese dataset, outperforming its performance on the English dataset. In total, AASIST2 demonstrates the best robustness when facing fade in & out. Logarithmic fade has the most significant impact on detectors' performance. When the proportion of logarithmic fade increases to 0.3, some detectors like RawGAT-ST tend to largely classify deepfake voices as real voices, while linear fade has the least impact. Detectors show noticeable performance differences across datasets of different languages.

5 Models vs. Humans

This section analyzes user study results, categorizing deepfake voices by deception levels (§ 3.3.2). We assess each detector's performance across these levels. Additionally, given large language models (LLMs)' versatility in various tasks [27], we explore an open-source multimodal LLM (MLLM)'s potential for deepfake voice identification.

Table 4: The results of the three user study experimental rounds. Results are consistent in both languages, indicating that participants' judgment abilities remain stable.

No.	English		Chinese	
	ACC	F1-Score	ACC	F1-Score
Exp.R1	59.28	62.30	78.20	80.06
Exp.R2	55.90	59.51	78.10	78.94
Exp.R3	54.92	60.01	77.50	79.01
Average	56.69	60.61	77.94	79.33

5.1 User Study Results and Deception Levels

Observing the overall performance of participants across the sample dataset, we compiled the statistics from the three user study experimental rounds and their average results, as shown in Table 4. For English, participants achieved average ACC of 56.69% and F1-Score of 60.61%. For Chinese, average ACC was 77.94% with F1-Score of 79.33%. Results were consistent across rounds, indicating stable human discernment of deepfake voices. Therefore, we can use the actual deception results of deepfake voices on users from the user study to grade the deepfake voices. Using the method in § 3.3, we created voice combinations and classified them into deception levels 0-2 based on human deception success. For English, 1,326 combinations were produced (210 level 0, 829 level 1, 287 level 2). For Chinese, 741 combinations were created (440 level 0, 184 level 1, 117 level 2). Table 5 displays participants' FARs for these three deception levels.

5.2 Performance Analysis by Deception Levels

We reassessed each detector's vulnerability to deception using the sample dataset mirroring the user study. We employed each detector's EER equal error point as the discrimination threshold (§ 4.2). For detectors with near-random AUC scores, we provide data performance without further discussion. Additionally, we evaluated `Qwen2-Audio` [13], a multimodal LLM (MLLM), to explore its potential in deepfake voice detection.

We analyzed detectors' FARs on voices graded by user study results, as shown in Table 5. A higher FAR indicates poorer performance. We found that for both Chinese and English datasets, on level 0 deepfake voices which are difficult to deceive humans, most detectors' FARs are higher than human performance. This means that deepfake voices easily identified as fake by humans are more likely to deceive the detectors, potentially misleading humans in practical applications. For the English dataset, we noticed that except for `RawBoost` and `RawGAT-ST`, detectors' FARs increased with deepfake voice levels, though not significantly. For level 1 and level 2 deepfake voices, the detectors' performance was notably better than humans'. For the Chinese dataset, all detectors showed a significant FAR increase from level 0 to 1, but less pronounced from level 1 to 2. Notably, `AASIST2`, `RawBMamba`, `RawGAT-ST`, and `RawNet2-Vocoder` performed better on level 2 than level 1. At levels 1 and 2, all detectors outperformed humans, but at level 1, most detectors' FARs (40%-50%) were not significantly better than human performance, except `AASITS2` (27.17% FAR). In summary, most detectors performed poorly on deepfake voices that are difficult to deceive humans. While they provide some assistance in detecting deepfake voices that humans find difficult to judge, the best FAR achieved was just 25.64%.

The MLLM, `Qwen-audio2`, exhibits relatively consistent FARs across various deception levels in both languages (Table 5). However, its performance is poor, with an F1-score of only 33.16% on the Chinese dataset and 0% on the English dataset. Interestingly, while MLLMs demonstrate superior capabilities in voice analysis, they struggle with detection tasks. This gap reveals that analysis proficiency doesn't guarantee deepfake voice detection capability.

5.3 Focus Analysis

Factors Affecting Human Judgment. Based on the participants' ratings of the quality of deepfake voices in the questionnaire (see § 3.3.2), we analyze the influencing factors chosen by humans when judging deepfake voices of varying deception levels. Figure 8 in the Appendix C shows the word clouds created using TF-IDF weights for the influencing factors at different levels. We found that the influencing factors provided by participants did not vary significantly across different scores. High-frequency factors like speech rate, emotion, pauses, and breathing were consistently highlighted in both English and Chinese datasets, while background noise, volume, interjections, and laughter received less attention. Timbre was also frequently mentioned as a key attribute. The timbre attribute is usually associated with specific speakers. In the voice collection process (see § 3.1), it cannot be guaranteed that all timbres are unfamiliar to the participants. Participants may have encountered some of these timbres on social media or through other channels, and this familiarity could potentially help them in judging deepfake voices.

Divergence in Feature Recognition: Models vs. Humans. Detectors typically use specific feature extractors (such as `OC-Softmax`) or data-driven deep learning models (such as `AASIST2`, which uses wav2vec 2.0) to learn features that distinguish deepfake voices from real voices. However, the features learned by these models do not always effectively reflect attributes that humans focus on, such as emotions. Deepfake voices that are easily recognizable by humans often exhibit noticeable differences from real voices in these human-perceived attributes. However, unless the detector's feature extraction can capture these attributes (e.g., the wav2vec 2.0 is widely applied in emotion recognition tasks), the detector may struggle to distinguish such deepfake voices from real voices, as these deepfake voices might resemble the average performance of voice features the model focuses on. Conversely, for high-quality deepfake voices that are difficult for humans to identify, although they may be similar to real voices in human-perceived attributes, the model's feature extractor might capture subtler differences than human perception. This may explain the interesting phenomenon in § 5.2: most detectors perform poorly when handling low-quality deepfake voices that are easily identifiable by humans but tend to outperform humans when dealing with high-quality deepfake voices that are challenging to discern.

6 Impact of Optimization Strategies

Data augmentation and cross-domain training are commonly used optimization strategies to improve the practical performance of detectors [68]. In this section, to investigate whether existing optimization strategies can help detectors address the potential risks identified by `VoiceWukong`, we evaluate detectors by applying RawBoost [53], a widely-adopted data augmentation method in the field, and the latest data augmentation method Targeted Augmented Data [5]. We also conduct cross-domain training using the latest cross-domain dataset CD-ADD [36] for the corresponding assessment.

6.1 Experimental setup

We selected `AASIST2`, the best-performing detector according to § 4, and `Res-TSSDNet`, one of the under-performing detectors, for evaluation. We conducted three experimental sets, the first two sets examined the effectiveness of data augmentation methods—RawBoost and Targeted Augmented Data—on

Table 5: FARs (%) of humans and various detectors on different levels of deepfake voices. Rows 2-4 and 9-11 of the table show the results on the English dataset, while the other rows present the results on the Chinese dataset.

	Human	AASIST	RawNet2	RawBoost	OC-Softmax	RawGAT-ST	CLAD
EN-Level0	18.97	28.33	25.71	23.81	30.48	24.04	17.62
EN-Level1	51.67	28.65	26.48	26.54	32.31	26.90	22.56
EN-Level2	82.64	30.14	29.27	24.74	35.54	26.13	24.39
CN-Level0	4.20	57.84	29.20	14.20	22.72	33.18	33.75
CN-Level1	50.54	47.01	46.20	47.28	40.76	47.28	41.03
CN-Level2	87.61	33.76	48.29	47.44	48.72	45.73	34.62

	Qwen-audio2	Res-TSSDNet	RawNet2-Vocoder	AASIST2	Raw PC-DARTS	RawBmamba	SAMO
EN-Level0	31.67	42.86	26.43	9.52	27.62	26.19	24.76
EN-Level1	32.15	51.69	29.07	14.60	28.59	29.07	27.20
EN-Level2	32.06	54.36	30.14	18.29	32.75	29.97	28.92
CN-Level0	34.87	48.30	29.32	9.55	21.82	30.91	56.02
CN-Level1	28.80	48.91	50.82	27.17	44.84	44.02	45.92
CN-Level2	28.63	45.30	48.29	25.64	49.57	35.90	33.33

both AASIST2 and Res-TSSDNet (based on ASVspoof2019). The third set investigated the benefits of multi-domain training by combining the CD-ADD dataset with ASVspoof2019. Since the original AASIST2 model was pre-trained with RawBoost, we established a non-augmented version as our baseline. All hyperparameters were maintained as specified in the original papers. In total, we completed six training and evaluation cycles across three experimental sets.

Table 6: EERs of detectors on `VoiceWukong` after applying different optimization strategies. "RB" represents RawBoost, "TA" represents Targeted Augmented Data, and "MD" represents Multi-Domain. "✓" means the method is applied, "×" means not applied.

Detector	RB	TA	MD	EN-EER(%)	ZH-EER(%)
AASIST2	×	×	×	31.40	32.67
	✓	×	×	13.50	13.54
	×	✓	×	29.39	35.22
	×	×	✓	48.18	49.89
Res-TSSDNet	×	×	×	50.06	49.91
	✓	×	×	50.02	50.06
	×	✓	×	50.03	49.90
	×	×	✓	50.08	49.79

6.2 Results with Optimization Strategies

Table 6 shows the EERs for all optimized detectors. In our analysis, baselines refer to the original detectors without augmentation or multi-domain training.

RawBoost. For `AASIST2`, the detector's performance on both Chinese and English datasets significantly improved with RawBoost data augmentation compared to the baseline. However, even after augmentation, Res-TSSDNet's performance remained similar to the baseline, nearly equivalent to random prediction. While RawBoost nearly enhanced detection across all manipulation types in `AASIST2`, the addition of Gaussian noise slightly increased the FARs in both datasets. In contrast, we found no improvement in Res-TSSDNet's performance for any specific manipulation with RawBoost.

Targeted Augmented Data. After applying Targeted Augmented Data, `AASIST2` saw a 2% increase in EER on the English dataset, while its performance on the Chinese dataset actually decreased. In the English dataset, `AASIST2` only showed performance improvements for specific manipulations. For instance, changing the sampling rate to 32K or 44.1K significantly reduced both FAR and FRR compared to the baseline. However, in most other cases, a decrease in either FAR or FRR was accompanied by an increase in the other. The most notable example was with added Gaussian noise, where `AASIST2` often misclassified voices as fake, resulting in a low FAR but a high FRR. In contrast, `Res-TSSDNet` did not exhibit significant performance improvements after data augmentation with this method, as indicated by the EER.

Multi-domain. Unfortunately, multi-domain training didn't improve the performance of `AASIST2` or `Res-TSSDNet` on `VoiceWukong`. In fact, after multi-domain training, `AASIST2` performed significantly worse than the baseline.

Although the RawBoost method significantly improved `AASIST2`, a considerable gap remains compared to the best performance reported in the original paper (0.82% EER). Additionally, the Targeted Augmentation Data and multi-domain training approaches did not substantially enhance the detectors' performance. **This indicates that our dataset continues**

to reveal potential risks that current detectors and optimization methods have not accounted for.

7 Discussion

7.1 Threats to Validity

Overlap Algorithms. During voice collection (see § 3.1.1), we are unaware of the underlying algorithms used by different commercial tools. As a result, their algorithms may overlap with the open-source models we identified. Without public disclosure, avoiding this overlap is challenging. Fortunately, our experimental results do not indicate significant issues.

Different Sample Rates. Our dataset (including subsets under resample manipulation) has a different sample rate (see § 3.1.2) from the training dataset, which could lead to significant performance differences for sampling rate-sensitive detectors compared to the results reported in the original paper, thereby affecting our evaluation.

Different Voice Duration. Differences in voice duration between our evaluation and training datasets (see § 3.2) could impact detector performance. However, in real-world scenarios, data duration is unpredictable and should not be an excuse for performance decline.

Number of Benign Voices. Our benign voice samples were sourced from two widely recognized datasets in the field: MAGICDATA and VCTK. Unfortunately, these datasets contain a limited number of benign voices and do not fully capture the diversity of accents in their respective languages. While such dataset limitations could potentially affect the evaluation of detector performance, we mitigated this concern by ensuring all detectors were trained on the identical training dataset, thus minimizing any bias from these limitations.

Participants' Random Choices. Despite adding attention tests to our user study (as mentioned in § 3.3.2), it is still impossible to completely prevent participants from making random choices for certain deepfake voices, which could affect the grading of deepfake voices. However, our statistical analysis of the F1-Scores and ACCs (see § 5.1) across three experimental rounds showed that the results were very consistent, indicating that our user study effectively reflected the participants' judgment ability.

7.2 Diverse Data Challenges

Datasets with a wide range of generation methods are crucial. Existing deepfake voice datasets, such as ASVspoof2021, excel in this by including voices generated by hundreds of algorithms. However, they share a common issue: a bias towards a single language and a lack of consideration for real-world voice manipulations. Our evaluation shows that detectors trained on a single language suffer significant performance drops with languages outside their training set and degrade further when facing certain manipulations. This suggests that detectors performing well on specific datasets may struggle with real-world deepfake voice detection. Although our work considers a variety of manipulations, it cannot cover all potential voice variations in practical scenarios. Recognizing the limitation of current datasets that focus on a single language, we made our best efforts to expand our work to include both Chinese and English. However, this still represents a significant limitation, preventing us from evaluating detector performance across a broader range of languages, which remains an important area for future research. In an era where people with diverse native languages can easily communicate and where generating deepfake voices is low-cost, it is crucial to consider the diversity of deepfake voice datasets. Developing detection methods that do not rely on language-specific features and enhancing detector robustness against various manipulations are vital for real-world deployment.

7.3 Human Focus: Key to Deepfake Detection

We categorized deepfake voices into three levels based on their ability to deceive humans, aiming to differentiate their quality. Results on the English dataset show that detector performance declines as the deception level increases, though the overall decline is not significant. Interestingly, detectors are more easily fooled by deepfake voices with low deception levels for humans, while they perform better on those with high deception levels. This suggests that when humans can accurately judge deepfake voices, relying solely on detectors may yield misleading results. Human judgment of deepfake voices is a complex process. Although we did not explore this process in detail, we asked participants to identify factors influencing their decisions. **The findings reveal common focus areas—such as emotion, speech rate, and pauses—when humans evaluate deepfake voices.** The ultimate goal of detectors is to help humans avoid deception. Future research should explore how to better align detector performance with human judgment, whether incorporating common human judgment features can reduce detectors' errors for low-deception deepfakes, and how to enhance overall detector effectiveness.

8 Conclusion

Our benchmark `VoiceWukong` fills the current gap in systematic and intuitive evaluation for deepfake voice detection. We collected a large set of deepfake voices in English and Chinese using a wide range of commercial tools and advanced open-source methods. Through manipulation, we addressed the limitations of existing datasets that are restricted to a single language or lack variation. We demonstrated the performance differences of various detectors under different manipulations, revealing that even the most advanced detectors still face

significant challenges in real-world applications. We also conducted a large-scale user study and found that detectors might mislead humans when dealing with deepfake voices of low deception, and identified the key features humans rely on to distinguish deepfake voices.

Ethics Considerations

Ethics. As mentioned in § 3.1.1, for each commercial tool, we carefully examined their terms of service to ensure they can be used for research purposes. We paid the necessary usage fees and ensured that `VoiceWukong` is non-commercial, thereby protecting their intellectual property. Additionally, to safeguard against potential misuse by dataset recipients, detailed usage restrictions will be further elaborated in our subsequent **Usage License**. Before recruiting participants, we obtained approval from our institution as an exempt study. We only collected participants' gender and age, without any identifiable or sensitive information, qualifying for exemption. As stated in § 3.3.2, our instructions clearly informed participants of the requirements and time limits, and they were free to withdraw at any time without restrictions.

Usage License. Our dataset is currently available exclusively to the academic research community through an application and approval process. To prevent misuse of the dataset or any potentially illegal activities, applicants must strictly comply with the following conditions before accessing our dataset:

1. Eligibility: Access to the dataset is limited to academic researchers for the purpose of evaluating detectors.
2. Redistribution Prohibition: Recipients are not permitted to redistribute the dataset without explicit permission.
3. Commercial Use Restrictions: The dataset may not be used for any commercial purposes, including but not limited to:
    - Product testing
    - Development activities
    - Commercial deployment
    - Model fine-tuning
    - Training commercial systems
    - Other profit-oriented uses
4. Legal Compliance: The use of the dataset for any activities prohibited by law is strictly forbidden.

Open Science

Artifacts Availability. Our artifacts including dataset, user study results, weighted models for experimental evaluation, and original outputs can be accessed through our permanent storage site (https://zenodo.org/records/13731918).

Acknowledgments

We are deeply grateful to our shepherd and anonymous reviewers for their insightful comments. This work was supported in part by the Key R&D Program of Hubei Province (2023BAB017 2023BAB079), HUST CSE-HongXin Joint Institute for Cyber Security, HUST CSE-FiberHome Joint Institute for Cyber Security, and the Xiaomi Young Talents Program. The full name of the authors' affiliation is Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology.

References

[1] The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/. Accessed: 2024-9-2.

[2] Fra.Txt details. https://www.kaggle.com/code/jasoncallaway/fra-txt-details, September 2018. Accessed: 2024-9-2.

[3] Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, August 2019. Accessed: 2024-6-7.

[4] Voicewukong leaderboard. https://voicewukong.github.io, 2024.

[5] Marcella Astrid, Enjie Ghorbel, and Djamila Aouada. Targeted augmented data for audio deepfake detection. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 346–350. IEEE, 2024.

[6] Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2033–2044, New York, NY, USA, 2023. Association for Computing Machinery.

[7] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[9] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech

corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE, 2017.

[10] Lian-Wu Chen, Wu Guo, and Li-Rong Dai. Speaker verification against synthetic speech. In *2010 7th International Symposium on Chinese Spoken Language Processing*, pages 309–312. IEEE, 2010.

[11] Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Lv Zhao, and Cunhang Fan. Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. *arXiv preprint arXiv:2406.06086*, 2024.

[12] Xingliang Cheng, Mingxing Xu, and Thomas Fang Zheng. Replay detection using cqt-based modified group delay feature and resnewt network in asvspoof 2019. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 540–545. IEEE, 2019.

[13] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[14] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8962–8966. IEEE, 2022.

[15] Phillip L De Leon, Bryan Stewart, and Junichi Yamagishi. Synthetic speech discrimination using pitch pattern statistics derived from image analysis. In *Proc. Interspeech: Portland, Oregon, USE*. 2012.

[16] Jingyi Deng, Chenhao Lin, Pengbin Hu, Chao Shen, Qian Wang, Qi Li, and Qiming Li. Towards benchmarking and evaluating deepfake detection. *IEEE Transactions on Dependable and Secure Computing*, pages 1–16, 2024.

[17] Siwen Ding, You Zhang, and Zhiyao Duan. Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[18] Thien-Phuc Doan, Long Nguyen-Vu, Souhwan Jung, and Kihun Hong. Bts-e: Audio deepfake detection using breathing-talking-silence encoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[19] Jonathan Driedger and Meinard Müller. A review of time-scale modification of music signals. *Applied Sciences*, 6(2):57, 2016.

[20] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

[21] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*, 2021.

[22] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

[23] Wanying Ge, Michele Panariello, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Partially-connected differentiable architecture search for deepfake and spoofing detection. *arXiv preprint arXiv:2104.03123*, 2021.

[24] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Raw differentiable architecture search for speech deepfake and spoofing detection. *arXiv preprint arXiv:2107.12212*, 2021.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Andrew Hery, Oluwaseyi Joseph, Olaoye Femi, and Hivez Luz. Audio deepfakes: threats to voice assistants and voice-activated systems. 2024.

[27] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *arXiv preprint arXiv:2308.10620*, 2023.

[28] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters*, 28:1265–1269, 2021.

[29] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In

ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6367–6371, 2022.

[30] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu. Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *Proc. Interspeech*, pages 3583–3587, 2020.

[31] Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. Sasv 2022: The first spoofing-aware speaker verification challenge. *arXiv preprint arXiv:2203.14732*, 2022.

[32] Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja. How deep are the fakes? focusing on audio deepfake: A survey. *arXiv preprint arXiv:2111.14203*, 2021.

[33] Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja. Audio deepfakes: A survey. *Frontiers in Big Data*, 5:1001063, 2023.

[34] Juntae Kim and Sung Min Ban. Phase-aware spoof speech detection based on res2net with phase network. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[35] Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1339–1349, January 2023.

[36] Yuang Li, Min Zhang, Mengxin Ren, Miaomiao Ma, Daimeng Wei, and Hao Yang. Cross-domain audio deepfake detection: Dataset and analysis. *arXiv preprint arXiv:2404.04904*, 2024.

[37] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[38] Lucian Lupşa-Tătaru. Novel technique of customizing the audio fade-out shape. *Applied Computer Science*, 14(3), 2018.

[39] Lucian Lupşa-Tătaru. Implementing the fade-in audio effect for real-time computing. *Applied Computer Science*, 15(2), 2019.

[40] Magic Data Technology Co., Ltd. Magicdata mandarin chinese read speech corpus, 05 2019.

[41] Juan M Martín-Doñas and Aitor Álvarez. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9241–9245. IEEE, 2022.

[42] Victoria Mingote, Antonio Miguel, Dayana Ribas, Alfonso Ortega Giménez, and Eduardo Lleida. Optimization of false acceptance/rejection rates and decision threshold for end-to-end text-dependent speaker verification systems. In *INTERSPEECH*, pages 2903–2907, 2019.

[43] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

[44] Jiahui Pan, Shuai Nie, Hui Zhang, Shulin He, Kanghao Zhang, Shan Liang, Xueliang Zhang, and Jianhua Tao. Speaker recognition-assisted robust audio deepfake detection. In *Interspeech*, pages 4202–4206, 2022.

[45] Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, Royi Nyameko, Srinivas Aluvala, and Vrince Vimal. Deepfake generation and detection: Case study and challenges. *IEEE Access*, 2023.

[46] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey, 2024.

[47] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[48] Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10. IEEE, 2019.

[49] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.

[50] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[51] Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 904–912, 2023.

[52] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710*, 2021.

[53] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE, 2022.

[54] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.

[55] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*, 2022.

[56] Hemlata Tak, Jee weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Graph attention networks for anti-spoofing, 2021.

[57] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Spoofing detection from a feature representation perspective. In *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2119–2123. IEEE, 2016.

[58] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[59] Anton Tomilov, Aleksei Svishchev, Marina Volkova, Artem Chirkovskiy, Alexander Kondratev, and Galina Lavrentyeva. Stc antispoofing systems for the asvspoof2021 challenge. In *Proc. ASVspoof 2021 Workshop*, pages 61–67, 2021.

[60] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2016.

[61] Jesús Villalba and Eduardo Lleida. Speaker verification performance degradation against spoofing and tampering attacks. In *FALA workshop*, pages 131–134, 2010.

[62] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2062–2070. IEEE, 2019.

[63] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1207–1216, New York, NY, USA, 2020. Association for Computing Machinery.

[64] Xin Wang and Junich Yamagishi. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. *arXiv preprint arXiv:2103.11326*, 2021.

[65] Zhenyu Wang and John HL Hansen. Audio anti-spoofing using a simple attention module and joint optimization based on additive angular margin loss and meta-learning. *arXiv preprint arXiv:2211.09898*, 2022.

[66] Thomas Wilmering, David Moffat, Alessia Milo, and Mark B Sandler. A history of audio effects. *Applied Sciences*, 10(3):791, 2020.

[67] Haolin Wu, Jing Chen, Ruiying Du, Cong Wu, Kun He, Xingcan Shang, Hao Ren, and Guowen Xu. Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning. *arXiv preprint arXiv:2404.15854*, 2024.

[68] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[69] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE, 2022.

[70] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*, 2023.

[71] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.

[72] Yongyi Zang, Jiatong Shi, You Zhang, Ryuichi Yamamoto, Jionghao Han, Yuxun Tang, Shengyuan Xu, Wenxiao Zhao, Jing Guo, Tomoki Toda, and Zhiyao Duan. Ctrsvdd: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection, 2024.

[73] XiaoHui Zhang, Jiangyan Yi, Chenglong Wang, Chu Yuan Zhang, Siding Zeng, and Jianhua Tao. What to remember: Self-adaptive continual learning for audio deepfake detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19569–19577, Mar. 2024.

[74] You Zhang, Fei Jiang, and Zhiyao Duan. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941, 2021.

APPENDIX

A Participants

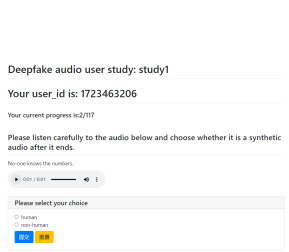
All participants were at least 18 years old and had a minimum of an undergraduate degree. We ensured that each participant was fluent in either English or Chinese to prevent language unfamiliarity from affecting their judgment of deepfake voices. Each participant received $5 for their participation. A total of 318 participants successfully completed the questionnaire within the specified time limit: 114 completed the Chinese questionnaire and 204 completed the English version. The participant pool had an average age of 22.40 years, comprising 64.47% males and 35.53% females. Detailed information about the participants for both Chinese and English questionnaires is shown in Table 7.

Table 7: Details of participants grouped by English and Chinese language. The English group's average age was 22.15, with a range of 18-29, while the Chinese group's average was 22.84, ranging from 20-29.

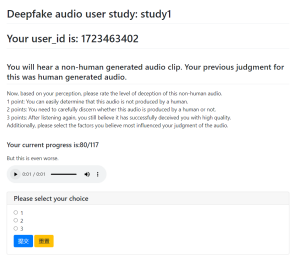
English		Chinese	
Sample size	204	**Sample size**	114
Age$_{avg}$(SD)	22.15 (1.89)	**Age$_{avg}$(SD)**	22.84 (2.93)
Age$_{max}$	29	**Age$_{max}$**	29
Age$_{min}$	18	**Age$_{min}$**	20
Female (%)	34.8%	**Female (%)**	36.84%
Male (%)	65.20%	**Male (%)**	63.16%
Non-binary (%)	0	**Non-binary (%)**	0

B Tasks for the Questionnaire

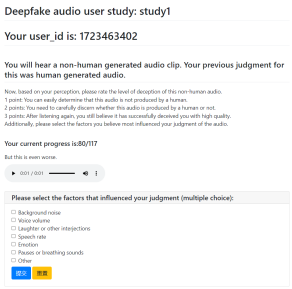
The Figure 5 shows examples from the English questionnaire. The tasks in the Chinese questionnaire are the same, only the language changed.



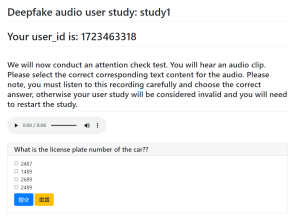
(a) Task 1: Whether the voice is human-generated?



(b) Task2-1: Rate the generation quality of the deepfake voice.



(c) Task2-2:What are the factors that influenced your judgment?



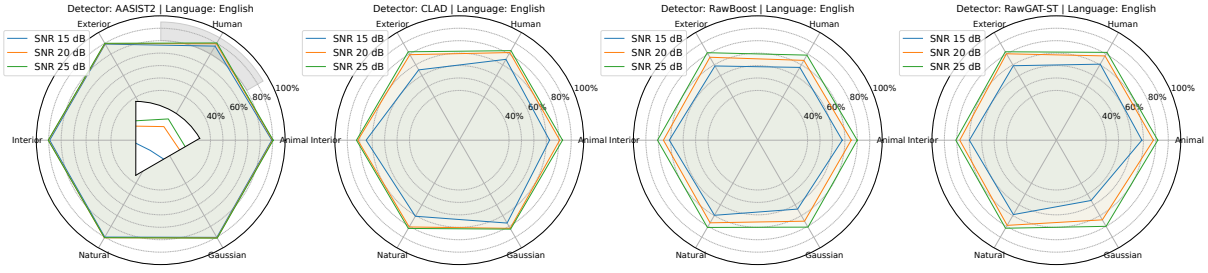
(d) Task 3: Attention test.

Figure 5: Examples of the three tasks in the questionnaire.

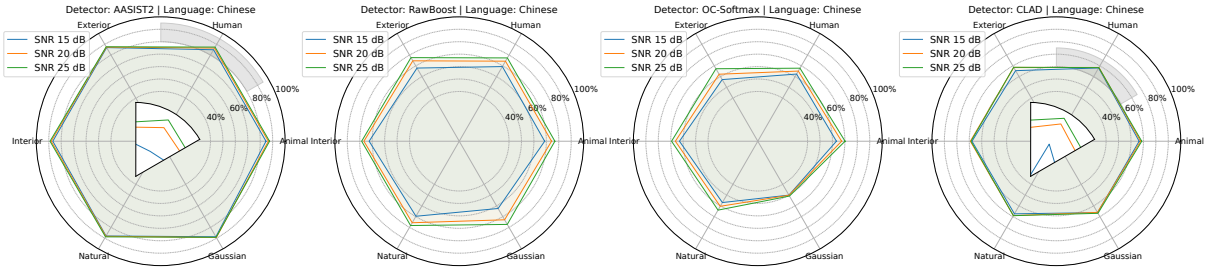
C The Results of Evaluation and Analysis

As mentioned in § 4.3, Figure 6 shows the ACCs of the top four AUC-ranked detectors under noise injection at different SNR levels. And Figure 7 shows spectrograms of the same voice at different volume levels and the cosine similarity between these spectrograms. The spectrograms' similarity of voices with different volume levels remains extremely high.

Figure 8 shows word clouds of influencing factors given by participants, weighted by TF-IDF. Subplots (a), (b), and (c) show results from the English dataset, while the remaining subplots present results from the Chinese dataset (see § 5.3).



(a) ACCs of the top four AUC-ranked detectors under noise injection at different SNR levels on the English dataset.



(b) ACCs of the top four AUC-ranked detectors under noise injection at different SNR levels on the Chinese dataset.

Figure 6: ACCs of the top four AUC-ranked detectors under noise injection at different SNR levels.

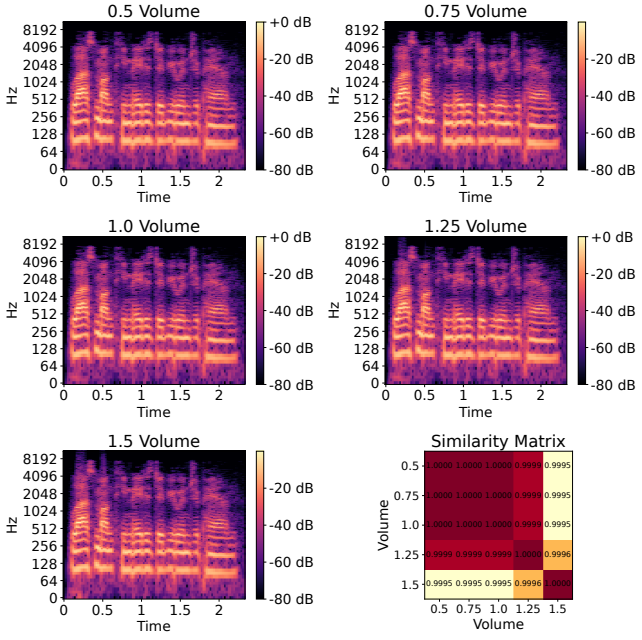


Figure 7: Spectrograms of the same voice at different volume levels and the cosine similarity between these spectrograms.

Figure 8: Word clouds of influencing factors given by participants, weighted by TF-IDF.