PC-Mix: Partial-Component Audio Spoofing Detection under Mixed Speech and Environmental Sound Conditions

Zhenshan Zhang[1], Xueping Zhang[1], Linxi Li[3], Yechen Wang[3], Ming Li[2,*]
[1]Digital Innovation Research Center, Duke Kunshan University, Kunshan, China
[2]School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
[3]OfSpectrum, Inc., Los Angeles, USA
[*]Corresponding author, mingli369@cuhk.edu.cn

***Abstract*—Recent studies on partial audio spoofing mainly focus on studio-recorded speech with temporal localization of spoofed segments. However, these studies often overlook realistic conditions where spoofed and bonafide segments simultaneously coexist across speech and environmental sound components. In this paper, we present PC-Mix, the first dataset for partial-component spoofing detection, where either or both audio components may be partially spoofed. In PC-Mix, bonafide and partially spoofed environmental-sound components are first constructed and mixed with speech signals from an existing partial-spoof dataset, producing audio in which either or both components may be locally manipulated. This design addresses two major gaps in existing partial spoofing benchmarks: the lack of realistic environmental sounds in speech partial spoofing scenarios and the absence of partial spoofing detection for environmental sound components. We establish standardized evaluation protocols and design a joint learning framework to optimize spoofing detection across speech, environmental sound, and mixed audio. Experiments highlight the increased difficulty introduced by mixed conditions. The results demonstrate that training under matched target conditions is more effective than directly transferring models trained on speech or environmental sound components.**

***Index Terms*—Audio Anti-Spoofing, Partial Spoofing Detection, Component-Level Audio Anti-Spoofing, Joint Learning**

I. Introduction

Audio anti-spoofing has become an important task in trustworthy speech and audio processing, as increasingly realistic synthetic and manipulated audio continues to pose serious security risks to real-world applications [1]–[3].

Spoofed audio can be generated or manipulated using increasingly powerful techniques, including voice conversion methods [4], [5], zero-shot speech synthesis methods [6], [7], and audio editing methods [8]–[10], posing significant threats to authentication systems and downstream applications.

Most research [11]–[14] only focuses on utterance-level spoof, ignoring localized manipulation within an utterance and treating the entire signal as either fully real or fully spoofed. More recently, attention has shifted toward partial spoofing, where only localized temporal regions of an utterance are manipulated while the rest remains bonafide [15], [16]. This scenario is more realistic, as modern audio editing tools enable fine-grained manipulation of selected segments without altering the entire recording. Existing benchmarks [15], [17]–[21] for partial-spoofing and temporal deepfake localization have shown that localized attacks are particularly challenging to detect.

However, these studies predominantly focus on the speech component alone and typically assume clean or studio-recorded conditions, overlooking the role of environmental background sounds [11], [12], [22], [23]. In real-world scenarios, audio recordings are inherently composed of multiple acoustic components, including foreground speech and environmental sounds. Recent work has begun to explore component-level spoofing, where spoofing may target either the speech component, the environmental component, or both [24], [25]. This formulation reflects practical scenarios where adversaries may manipulate only selected audio components such as speech or environmental sound or possibly both. Such mixed conditions introduce additional complexity, as detection systems must reason about multiple interacting sources rather than a single homogeneous signal.

Nevertheless, existing studies mainly consider component-level manipulation at the utterance level and do not explicitly address temporally localized spoofing within each component. To address this gap, we establish a partial component spoof evaluation protocol. We propose PC-Mix, a new dataset for partial-component spoofing under mixed speech and environmental sound conditions. In PC-Mix, we construct bonafide and partially spoofed environmental-sound components and mix them with speech signals from an existing partial-spoof dataset. Each resulting sample may contain localized spoofing in the environmental sound component, the speech component, or both. Original recordings are also included to distinguish naturally recorded audio from constructed mixtures. A joint learning framework is further proposed for frame-level partial-component spoofing detection. The framework adopts a two-stage training strategy, first learning component-specific cues for speech spoofing, environmental-sound spoofing, and original-mix discrimination, and then jointly optimizing them under a unified partial-component spoofing objective. This design combines component-level localization with mix-level scene discrimination.

Our main contributions are summarized as follows:

- We introduce partial-component spoofing detection as a new evaluation task for partially edited audio, where localized spoofing may occur in speech, environmental sound, or both within a mixed acoustic scene.
- We propose PC-Mix, the first dataset for partial-component spoofing detection and the first benchmark dataset that includes partially spoofed environmental sounds. PC-Mix addresses two major gaps in existing partial spoofing benchmarks: the lack of realistic environmental sounds in speech partial spoofing scenarios and the absence of partial spoofing detection for environmental sound components.
- We propose a jointly trained baseline framework for partial-component spoofing detection. The framework performs speech-component spoofing detection, environmental-sound spoofing detection, and original-mix discrimination within a unified model.
- The source code and dataset are publicly released[1] to facilitate reproducible research and future studies.

II. PC-Mix Dataset

PC-Mix is designed to study partial-component spoofing under mixed speech and environmental sound conditions. It contains two types of data: original samples and constructed mixed samples. The original samples are real-world recordings that naturally contain speech and environmental sounds. The constructed mixed samples are generated by mixing partially spoofed speech utterances with partially spoofed environmental sound components.

A. Audio Source

1) Original recorded audio: Original recordings are included as a reference domain for real speech–environmental sound scenes. These samples are not created through artificial mixing, representing real acoustic conditions where speech and environmental sounds coexist. In PC-Mix, original recordings are collected from VGGSound [26] and SBCASE [27]. They are used to help distinguish naturally recorded audio from constructed speech–environmental sound mixtures.

2) Partially spoofed speech: For the speech component, audio is from PartialSpoof [15].

3) Partially spoofed environmental sound: This work introduces the first environmental-sound partial spoofing dataset designed for component-level spoofing analysis. For the environmental sound component, partially spoofed samples are constructed by inserting real or generated event sounds into real environmental background recordings.

In the training set, SONYC [28] is used as the background source. Spoofed event sounds are generated by AudioLDM2 [29] and real events are curated from UrbanSound8K [30]. The event sounds are inserted into background recordings using a controlled ducking-overlay strategy [31], where the background is locally attenuated during the inserted event. This preserves acoustic overlap while keeping the event perceptually salient.

In the evaluation set, five evaluation subsets are constructed, denoted as E0–E4. E0 serves as the in-domain evaluation set. It uses SONYC backgrounds, together with a new event pool composed of non-overlapping events generated by AudioLDM2 and real events from FSD50K [32]. E1 evaluates generator shift by replacing AudioLDM 2 with AudioGen [33]. E2 evaluates fusion shift by replacing the default ducking-overlay strategy with hard replacement followed by crossfade [34], while keeping the source domains unchanged. E3 evaluates background-domain shift by replacing SONYC with DEMAND [35]. E4 evaluates noise-domain shift by using WHAM! noise as the background source [36], resulting in more noise-dominated acoustic conditions. The statistics of the training and evaluation splits are summarized in Table I.

TABLE I
STATISTICS OF THE ENVIRONMENTAL PARTIAL-SPOOF COMPONENTS.

Split	#All	#Bona	#Spoof	BG	Event Source	Fusion	Shift
Train	25,380	12,715	12,665	SONYC	ALDM2/US8K	Duck.	Train
E0	17,812	8,931	8,881	SONYC	ALDM2/FSD50K	Duck.	ID
E1	14,248	7,181	7,067	SONYC	AudioGen	Duck.	Gen. OOD
E2	14,248	7,045	7,203	SONYC	ALDM2	Cross.	Fus. OOD
E3	17,809	8,757	9,052	DEMAND	ALDM2	Duck.	BG OOD
E4	7,125	3,562	3,563	WHAM!	ALDM2	Duck.	Noise OOD

Note: BG denotes the background source. OOD denotes the out-of-domain setting. ID denotes the in-domain setting. Duck., Cross., ALDM2, and US8K denote ducking overlay, crossfade-based fusion, AudioLDM2, and UrbanSound8K, respectively.

TABLE II
PC-MIX CLASS DEFINITIONS FOR UTTERANCE-LEVEL AND FRAME-LEVEL EVALUATION.

Level	ID	Mixed	Speech	Env.	Description
Utt.	0	No	–	–	Original outdoor recording
	1	Yes	bonafide	bonafide	bonafide speech with bonafide env.
	2	Yes	Spoofed	bonafide	spoofed speech with bonafide env.
	3	Yes	bonafide	Spoofed	bonafide speech with spoofed env.
	4	Yes	Spoofed	Spoofed	spoofed speech with spoofed env.
Frame	0	Yes	bonafide	bonafide	bonafide speech and env. frame
	1	Yes	Spoofed	bonafide	spoofed speech frame
	2	Yes	bonafide	Spoofed	spoofed env. frame
	3	Yes	Spoofed	Spoofed	spoofed speech and env. frame

B. Mix Process

Each partially spoofed speech utterance is paired with a partially spoofed environmental-sound recording and inserted at a random position on its timeline, with the output duration kept equal to the environmental-sound component by zero-padding shorter speech or truncating longer speech. All audio is sampled at 16 kHz. Training and evaluation mixtures are generated with nominal speech-to-environmental-sound SNRs of 8 dB and 10 dB, respectively, using fixed speech and environmental-sound gains of 1.0 and 1.4 with clipping protection; the effective SNR may vary slightly because environmental-sound scaling is computed for each pair.

PC-Mix provides utterance-level and frame-level labels for both speech and environmental-sound components. Speech

[1]https://github.com/Alphawarheads/PC-Mix

TABLE III
UTTERANCE-LEVEL CLASS DISTRIBUTION OF PC-MIX.

Split	Orig	BB	BS	SB	SS	#Total	Duration.
Train	12,155	1,287	1,293	11,427	11,373	37,535	98.95h
Eval	17,809	3,703	3,666	31,773	32,100	89,051	41.71h
Total	29,964	4,990	4,959	43,200	43,473	126,586	140.65 h

Note: Orig denotes original recordings. For mixed samples, the first and second letters denote the speech and environmental-sound labels, respectively. B and S denote bonafide and spoofed.

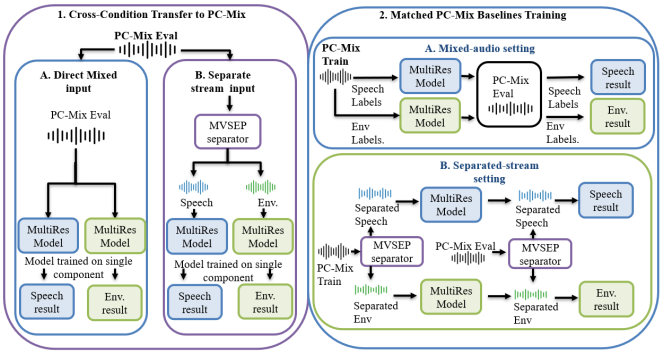


Fig. 1. Experimental pipeline for baseline studies. Single-component detectors and matched PC-Mix-trained models are evaluated on the PC-Mix evaluation set. Original recordings are excluded from component-level baselines

labels are inherited from PartialSpoof, while environmental-sound labels are derived from the constructed environmental-sound component. Segment labels are provided at 40, 80, 160, 320, and 640 ms resolutions. In addition, original-mix labels are also provided to distinguish naturally recorded audio from constructed mixtures. Based on these annotations, PC-Mix defines five utterance-level classes and four frame-level classes, as summarized in Table II; frame-level classes are defined only for constructed mixtures.

Overall, PC-Mix contains 126,586 clips and 140.65 hours of audio. The utterance-level class distribution of the training and evaluation splits is reported in Table III.

III. METHODOLOGY

A. Baseline System

We conduct two types of PC-Mix baseline evaluations, as illustrated in Fig. 1. Original real-world recordings in PC-Mix are excluded from these component-level baseline experiments because they do not provide separate speech- or environmental-sound spoofing annotations; they are later reintroduced in the joint learning experiments.

1) Cross-Condition Transfer to PC-Mix: We trained single-component speech and environmental-sound detectors on PartialSpoof and the constructed environmental-sound data, respectively. Then we transferred the single-component detectors to PC-Mix evaluation. During evaluation, two input settings are considered. In the mixed-audio input setting, both detectors are directly evaluated on the PC-Mix mixtures using their corresponding component labels. In the separated-stream input setting, each PC-Mix waveform is first separated into speech and environmental-sound streams using the MVSEP separator [37], after which each detector is evaluated on its corresponding separated stream.

2) Matched PC-Mix Baselines: We train matched PC-Mix detectors under two input settings. In the mixed-audio input setting, speech-focused and environmental-sound-focused detectors are trained on PC-Mix training mixtures and evaluated on PC-Mix evaluation mixtures using the corresponding component labels. In the separated-stream input setting, both the training and evaluation mixtures are first separated into speech and environmental-sound streams using MVSEP [37]; each detector is then trained and evaluated on its corresponding separated stream. These matched baselines provide references for assessing the effect of matched mixed-condition training and for comparing with the proposed joint learning framework.

B. Three-head Joint Learning Framework

1) Framework Architecture: A three-head joint learning framework is established for partial-component spoofing detection, as illustrated in Fig. 2. The framework consists of three MultiResoModel [18] branches: a speech branch, an environmental sound branch, and a mix branch. The speech and environmental sound branches are tailored to detect partial spoofing in the corresponding audio components, while the mix branch is designed to distinguish constructed speech–environmental sound mixtures from original real-world audio recordings. Each branch follows the same multi-resolution architecture, where SSL-based acoustic representations are processed by temporal modeling layers and multiple detection heads. During joint training, the three branches are optimized in a unified loss, allowing component-level spoofing cues and mix-level detection cues to be learned simultaneously.

Given an input waveform $\mathbf{x}$, the three branches produce:

$$\mathbf{z}^s = f_s(\mathbf{x}), \quad \mathbf{z}^e = f_e(\mathbf{x}), \quad \mathbf{z}^m = f_m(\mathbf{x}), \tag{1}$$

where $f_s(\cdot)$, $f_e(\cdot)$, and $f_m(\cdot)$ denote the speech, environmental sound, and mix branches, respectively. Their outputs $\mathbf{z}^s$, $\mathbf{z}^e$, and $\mathbf{z}^m$ contain multi-resolution frame-level prediction scores and an utterance-level prediction score. The speech and environmental sound branch predict partial spoofing in the corresponding components, while the mix branch predicts whether the input audio is an original recording or a constructed mix.

2) Two-Stage Training Strategy: The framework is trained in two stages. In the first stage, the three branches are trained independently using their corresponding labels. The speech branch is trained with speech-component partial spoofing labels, the environmental sound branch is trained with environmental-sound partial spoofing labels, and the mix branch is trained with original-mix labels. In the second stage, the three pretrained branches are combined into a single joint framework and optimized together. For the mix branch, only the utterance-level original-mix loss is used during second-stage joint training, while its frame-level outputs are excluded from loss supervision and not used for reported metrics. This design allows each branch to first learn component-

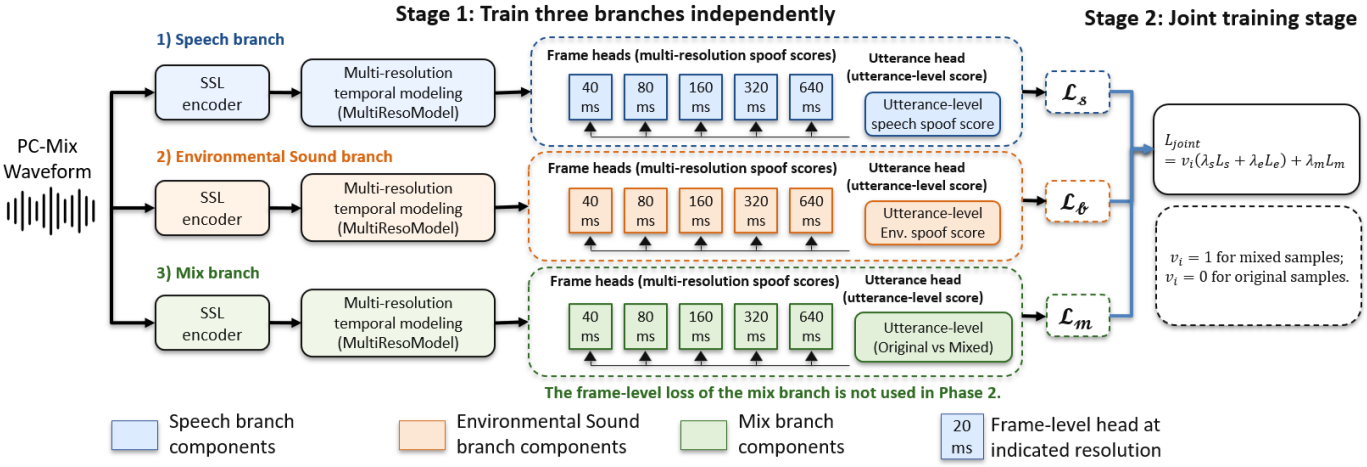


Fig. 2. Two-stage three-head joint learning framework for partial-component spoofing detection. Stage 1 independently trains speech, environmental-sound, and mix branches with their corresponding labels on PC-Mix train. Stage 2 jointly optimizes the pretrained branches, applying component-level losses only to constructed mixtures and utterance-level original-mix supervision to the mix branch. Env denotes Environmental sounds.

specific spoofing cues and then adapt jointly under the partial-component spoofing setting.

3) Joint Training: Let $\mathcal{K}$ denote the set of frame-level temporal resolutions. For i-th sample, the speech branch output consists of frame-level scores $\{\mathbf{z}^s_{i,k}\}_{k\in\mathcal{K}}$ and an utterance-level score $z^s_{i,u}$. Similarly, the environmental-sound branch produces $\{\mathbf{z}^e_{i,k}\}_{k\in\mathcal{K}}$ and $z^e_{i,u}$, and the mix branch produces $\{\mathbf{z}^m_{i,k}\}_{k\in\mathcal{K}}$ and $z^m_{i,u}$ respectively.

The corresponding speech and environmental-sound labels are denoted as $\{\mathbf{y}^s_{i,k}\}_{k\in\mathcal{K}}$, $y^s_{i,u}$, $\{\mathbf{y}^e_{i,k}\}_{k\in\mathcal{K}}$, and $y^e_{i,u}$. For the mix branch, $y^m_{i,u}$ denotes the utterance-level original-mix label, where $y^m_{i,u} = 1$ indicates a constructed mix and $y^m_{i,u} = 0$ indicates an original sample. During first-stage independent training of the mix branch, $\mathbf{y}^m_{i,k}$ denotes the corresponding frame-level original-mix label at resolution k. The P2SGrad loss [38] is used for all binary detection branches.

The speech-branch multi-resolution loss is defined as

$$\mathcal{L}^s_i = \sum_{k\in\mathcal{K}} \ell(\mathbf{z}^s_{i,k}, \mathbf{y}^s_{i,k}) + \ell(z^s_{i,u}, y^s_{i,u}). \tag{2}$$

The environmental sound branch follows the same form:

$$\mathcal{L}^e_i = \sum_{k\in\mathcal{K}} \ell(\mathbf{z}^e_{i,k}, \mathbf{y}^e_{i,k}) + \ell(z^e_{i,u}, y^e_{i,u}). \tag{3}$$

For the mix branch, first-stage independent training uses both frame-level and utterance-level supervision:

$$\mathcal{L}^{m,\mathrm{P1}}_i = \sum_{k\in\mathcal{K}} \ell(\mathbf{z}^m_{i,k}, \mathbf{y}^m_{i,k}) + \ell(z^m_{i,u}, y^m_{i,u}). \tag{4}$$

During second-stage joint training, only the utterance-level original-mix loss is used:

$$\mathcal{L}^{m,\mathrm{P2}}_i = \ell(z^m_{i,u}, y^m_{i,u}). \tag{5}$$

Since original samples do not contain separated speech and environmental-sound spoofing annotations, component-level losses are applied only to constructed mixtures. This is implemented as

$$v_i = \begin{cases} 1, & y^m_{i,u} = 1, \\ 0, & y^m_{i,u} = 0. \end{cases} \tag{6}$$

The final joint objective over a mini-batch $\mathcal{B}$ is

$$\mathcal{L}_{\text{joint}} = \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \left(\lambda_s v_i \mathcal{L}^s_i + \lambda_e v_i \mathcal{L}^e_i + \lambda_m \mathcal{L}^{m,\mathrm{P2}}_i \right). \tag{7}$$

where λ_s, λ_e, and λ_m are the loss weights for the speech, environmental sound, and mix branches. Thus, mix-level supervision is applied to all samples, while speech and environmental sound losses are masked out for original recordings without component-level annotations.

IV. Experiments

A. Experimental Settings

1) Data Processing: Audio samples are resampled to 16 kHz and processed as 9.6-second segments with a trailing padding duration of 0.005 s. All models use 160 ms frame-level supervision and produce frame-level outputs at the same resolution, together with an utterance-level output. Random seeking and mask-based training are enabled.

2) Training Configuration: All detectors are implemented with MultiResoModel [18] as the backbone. All models are optimized using Adam with a learning rate of 1×10^{-5}. A StepLR scheduler is applied with a step size of 10 epochs and a decay factor of 0.5.

Equal branch weights are used in all joint-training experiments, i.e., $\lambda_s = \lambda_e = \lambda_m = 1$. During joint training, the mix branch is trained on all samples, while the speech and environmental-sound branches are optimized only when valid component-level annotations are available.

TABLE IV
SINGLE-COMPONENT PARTIAL-SPOOFING SPEECH/ENV DETECTOR TRAINED AND EVALUATED ON SPEECH AND ENVIRONMENTAL-SOUND EVAL SET.

Train data	Eval data	Epoch	Utt. EER (%)	Frame EER (%)
Speech train	Speech eval	25	2.16	11.36
Env. sound train	Env. sound eval	25	3.55	5.89

TABLE V
ENVIRONMENTAL-SOUND PARTIAL-SPOOFING DETECTION PERFORMANCE ACROSS DIFFERENT IN-DOMAIN AND OUT-OF-DOMAIN EVALUATION SUBSETS.

Train data	Eval data	Epoch	Utt. EER (%)	Frame EER (%)
Env. sound train	Env. sound eval E0	25	0.67	2.47
Env. sound train	Env. sound eval E1	25	1.88	7.98
Env. sound train	Env. sound eval E2	25	0.53	2.57
Env. sound train	Env. sound eval E3	25	9.15	13.04
Env. sound train	Env. sound eval E4	25	3.69	3.93

3) Evaluation Metrics: Performance is evaluated using both binary and multi-class metrics. For the baseline evaluations, only equal error rate (EER) is reported at both the utterance and frame levels. For each binary branch in the joint framework, utterance-level metrics include EER, F1 score, and accuracy. For the speech and environmental sound branches, the same binary metrics are also computed at the frame level. Branch predictions are further mapped to the class definitions in Table II. At the utterance level, five-class classification is evaluated with a mix-gated decision rule. Accuracy, macro F1, and weighted F1 are reported for this evaluation. At the frame level, four-class classification is evaluated on mixed utterances based on the authenticity states of speech and environmental sounds, using the same multi-class metrics. Since original-mix supervision is defined at the recording level, the mix-branch frame-level outputs are not used in joint evaluation; therefore, frame-level five-class classification is not reported.

B. Experimental results

1) Single-Component Partial Spoofing Detection Results: We train MultiResoModel-based speech and environmental-sound spoofing detectors on PartialSpoof and the constructed partial-spoof environmental-sound training data, respectively, and evaluate them on the corresponding single-component test sets. Table IV shows that both speech and environmental-sound spoofing detectors perform well under matched conditions, indicating that each detection head is effective on its corresponding component.

Table V presents the evaluation results of the environmental-sound detector on the proposed E0–E4 subsets. The results show that these subsets expose different challenges for partial environmental-sound spoofing detection. The detector performs well on the in-domain E0 set and remains stable under fusion shift in E2, suggesting that inserted spoofed events can be localized when the background domain is consistent. However, performance degrades most under background-domain shift in E3, indicating that environmental context mismatch is a key challenge. These results demonstrate that the proposed subsets provide a useful benchmark for evaluating localized spoofing detection under different domain shifts.

TABLE VI
CROSS-CONDITION TRANSFER EVALUATION FROM THE SINGLE-COMPONENT SPEECH/ENV TRAINED DETECTOR TO PC-MIX MIXED-AUDIO AND SEPARATED SPEECH OR ENVIRONMENT SOUND INPUTS.

Setting	Target	Input	Epoch	Utt. EER (%)	Frame EER (%)
Mixed	Speech	Mixed audio	25	29.35	32.88
Mixed	Env. sound	Mixed audio	25	44.80	33.44
Separated	Speech	Sep. speech	25	51.38	37.51
Separated	Env. sound	Sep. env.	25	22.89	17.65

Note: Env. sound denotes environmental sound, and Utt. denotes utterance-level.

TABLE VII
MATCHED PC-MIX BASELINES FOR COMPONENT-SPECIFIC SPOOFING DETECTION, WHERE BASELINE MODELS ARE TRAINED AND EVALUATED USING EITHER THE COMPLETE PC-MIX TRAIN AUDIO OR THE SEPARATED SPEECH AND ENVIRONMENTAL-SOUND STREAMS.

Train & eval setting	Target	Epoch	Utt. EER (%)	Frame EER (%)
Mixed-audio input	Speech	30	8.53	22.52
Mixed-audio input	Env. sound	30	3.57	5.68
Separated-stream input	Speech	30	7.83	22.61
Separated-stream input	Env. sound	30	7.43	8.21

Note: Env. sound denotes environmental sound, and Utt. denotes utterance-level.

2) Baseline result: This experiment corresponds to the cross-condition transfer setting, where single-component detectors are evaluated directly under PC-Mix conditions. As shown in Table VI, both input settings suffer substantial degradation compared to the single-component results in Table IV. This indicates that detectors trained only on speech or environmental sounds cannot be directly transferred to PC-Mix conditions. Although separated-stream evaluation improves environmental-sound detection over mixed-audio evaluation, it remains far from the single-component setting and further degrades speech detection.

Table VII reports the matched PC-Mix baseline results, where component-specific detectors are trained and evaluated under the same PC-Mix input setting. Compared with cross-condition transfer in Table VI, matched PC-Mix training greatly improves performance. In the mixed-audio setting, speech utterance-level EER decreases from 29.35% to 8.53%, and environmental-sound utterance-level EER decreases from 44.80% to 3.57%. This confirms that the cross-condition degradation mainly comes from the mismatch between single-component training and mixed acoustic conditions. The separated-stream setting also improves over cross-condition transfer, especially for environmental-sound detection, where utterance-level EER decreases from 22.89% to 7.43%. However, it does not consistently outperform mixed-audio training: speech results are close across the two settings, while environmental-sound detection is clearly worse after separation. This suggests that source separation may introduce artifacts or remove environmental-sound spoofing cues. Therefore, the proposed joint learning framework adopts the

TABLE VIII

ABLATION COMPARISON BETWEEN THE ASSEMBLED INDEPENDENTLY TRAINED BRANCHES AND THE JOINTLY TRAINED THREE-HEAD MODEL. BOTH MODELS ARE TRAINED ON THE PC-MIX TRAINING SET AND EVALUATED ON THE PC-MIX EVALUATION SET. "PRE-JOINT" DENOTES THE ASSEMBLED MODEL WITHOUT SECOND-STAGE JOINT OPTIMIZATION, AND "JOINT-E8" DENOTES THE MODEL AFTER EIGHT EPOCHS OF JOINT TRAINING.

(a) Binary head performance				
Level	Head	EER ↓	F1 ↑	ACC ↑
Utterance	Speech	8.72 → 7.86 (-0.86)	94.56 → 94.58 (+0.02)	89.69 → 89.72 (+0.03)
	Env.	3.59 → 3.12 (-0.47)	76.41 → 92.67 (+16.26)	69.36 → 92.17 (+22.81)
	Mix	0.96 → 0.42 (-0.54)	99.43 → 99.22 (-0.21)	99.10 → 98.77 (-0.33)
Frame	Speech	22.54 → 21.84 (-0.70)	71.70 → 72.55 (+0.85)	77.27 → 78.07 (+0.80)
	Env.	5.18 → 3.87 (-1.31)	89.93 → 90.80 (+0.87)	96.94 → 96.96 (+0.02)
(b) Multi-class decision performance				
Task		ACC ↑	Macro F1 ↑	Weighted F1 ↑
Utterance 5-class		69.40 → 85.12 (+15.72)	44.80 → 54.31 (+9.51)	64.78 → 81.52 (+16.74)
Frame 4-class		74.92 → 75.72 (+0.80)	71.58 → 72.58 (+1.00)	75.16 → 76.04 (+0.88)

Note: In each entry, the value before → is the Pre-joint result and the value after "→" is the Joint-E8 result. Values in parentheses indicate absolute changes after joint training. The mix branch is evaluated only at the utterance level for original-mix discrimination. Env. denotes Environmental sound. All values are in %.

mixed-audio input setting rather than relying on explicit source separation models.

3) Ablation Study: Table VIII presents an ablation study of the second-stage joint optimization by comparing the jointly trained model with an assembled model composed of the three independently trained branches from stage 1. In this baseline, the independently trained speech and environmental-sound branches from the previous mixed-audio input experiment, together with the independently trained mix branch, are directly assembled and evaluated without second-stage joint optimization. Each entry is reported as "Pre-joint → Joint-E8", where Pre-joint denotes the assembled stage-one baseline and Joint-E8 denotes the model after eight epochs of joint training.

Results show that joint training is more effective than directly assembling independently trained heads. The improvement is most pronounced for the environmental-sound branch, whose utterance-level F1 and accuracy increase from 76.41% and 69.36% to 92.67% and 92.17%, respectively. This suggests that environmental-sound spoofing detection is particularly sensitive to the final mixed-audio training condition, where spoofing cues are entangled with speech and other acoustic components. In contrast, the speech branch shows smaller but consistent gains, indicating that it is already relatively stable after stage-one training and mainly benefits from refinement during joint optimization. The mix branch also maintains strong original-mix discrimination, with a slight decrease in EER, although its F1 and accuracy slightly decrease, suggesting that its operating point may require further calibration.

The multi-class results further confirm the benefit of joint training. At the utterance level, 5-class accuracy increases from 69.40% to 85.12%, and weighted F1 also improves substantially, indicating better consistency among the mix, speech, and environmental-sound branches. Macro F1 also increases, although it remains lower than weighted F1 due to the imbalanced class distribution, especially for the smaller mixed classes. This suggests that joint training improves overall decision consistency while partially mitigating the effect of class imbalance. At the frame level, the gains are smaller but still positive, with accuracy, macro F1, and weighted F1 all increasing after joint training. This indicates that frame-level localization remains more difficult than utterance-level classification, since the model must identify short spoofed regions under overlapping speech and environmental-sound components. Overall, joint training mainly strengthens utterance-level component reasoning, while fine-grained temporal localization remains challenging under mixed-audio conditions.

V. CONCLUSION

In this paper, we introduce partial-component spoofing detection as a new task for partially edited audio. Here, spoofing may occur in speech, environmental sounds, or both within a mixed acoustic scene. We thus propose PC-Mix, a benchmark for this task. It is, to the best of our knowledge, the first dataset that includes partially spoofed environmental sounds. It addresses two key gaps in prior benchmarks: the lack of realistic environmental sounds in speech-based partial spoofing and the lack of evaluation for environmental sound spoofing. We further propose a jointly trained baseline framework for this task. It performs speech-component spoofing detection, environmental-sound spoofing detection, and original-mix discrimination within a unified model.

ACKNOWLEDGMENT

The authors used OpenAI ChatGPT only for language editing, grammar checking, and clarity improvement across the manuscript. The tool was not used to generate research ideas, experimental designs, datasets, experimental results, figures, tables, or conclusions. The authors reviewed and verified all content, claims, and results, and remain fully responsible for the manuscript.

REFERENCES

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and Countermeasures for Speaker Verification: A Survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[2] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[3] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[4] K. Wang, W. Guan, Z. Jiang, H. Huang, P. Chen, W. Wu, Q. Hong, and L. Li, "Discl-VC: Disentangled Discrete Tokens and In-Context Learning for Controllable Zero-Shot Voice Conversion," in *Proceedings of Interspeech*, 2025, pp. 1383–1387.

[5] P. Ren, W. Guan, K. Wang, P. Chen, Q. Hong, and L. Li, "ReFlow-VC: Zero-shot Voice Conversion Based on Rectified Flow and Speaker Feature Optimization," in *Proceedings of Interspeech*, 2025, pp. 1388–1392.

[6] Y. Yang, S. Liu, J. Li, Y. Hu, H. Wu, H. Wang, J. Yu, L. Meng, H. Sun, Y. Liu, Y. Lu, K. Yu, and X. Chen, "Pseudo-Autoregressive Neural Codec Language Models for Efficient Zero-Shot Text-to-Speech Synthesis," in *Proceedings of the ACM International Conference on Multimedia*, 2025, pp. 9316–9325.

[7] X. Li, J. Xing, X. Xing, Z. Li, and X. Xu, "SA-RAS: Speaker-Aware Style Retrieval Augmented Generation for Expressive Zero-Shot Text-to-Speech Synthesis," in *Proceedings of Interspeech*, 2025, pp. 4388–4392.

[8] S.-F. Huang, H.-C. Kuo, Z. Chen, X. Yang, P.-J. Ku, A. Jukic, H. Yang, Y. Tsao, Y.-C. F. Wang, H. yi Lee, and S.-W. Fu, "VoiceNoNG: Robust High-Quality Speech Editing Model without Hallucinations," in *Proceedings of Interspeech*, 2025, pp. 3469–3473.

[9] M. Ungersböck, F. Grötschla, L. A. Lanzendörfer, J. Y. Yi, C. Choi, and R. Wattenhofer, "SAO-Instruct: Free-form Audio Editing using Natural Language Instructions," in *Advances in Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=EbjFkpXMke

[10] Z. Lan, Y. Hao, and M. Zhao, "SmartDJ: Declarative Audio Editing with Audio Language Model," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. [Online]. Available: https://openreview.net/forum?id=eNmANCkefl

[11] X. Zhang, Z. Zhang, Y. Wang, L. Li, L. Jin, and M. Li, "MultiAPI Spoof: A Multi-API Dataset and Local-Attention Network for Speech Anti-Spoofing Detection," in *Proceedings of Interspeech*, 2026, accepted.

[12] Z. Zhang, X. Zhang, Y. Wang, L. Jin, and M. Li, "The Impact of Audio Watermarking on Audio Anti-Spoofing Countermeasures," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 16 237–16 241.

[13] L. Buera, H. Delgado, D. Colibro, and A. Miguel, "Combining SSL Speech Features, Contextual Transformers and Mamba Models for Realistic Audio Spoofing Detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[14] D.-T. Truong, T. Liu, J. Li, R. Tao, K. A. Lee, and E. S. Chng, "Addressing Gradient Misalignment in Data-Augmented Training for Robust Speech Deepfake Detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[15] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023.

[16] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, X. Zhang, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, H. Li, Z. Lian, and B. Liu, "ADD 2022: The First Audio Deep Synthesis Detection Challenge," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9216–9220.

[17] Y. Zhang, B. Tian, L. Zhang, and Z. Duan, "PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing," in *Proceedings of Interspeech*, 2025, pp. 5353–5357.

[18] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng, "LlamaPartialSpoof: An LLM-Driven Fake Speech Dataset Simulating Disinformation Generation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[19] O. Baser, A. E. Tanriverdi, S. Vishwanath, and S. Chinchali, "PhonemeFake: Redefining Deepfake Realism with Language-Driven Segmental Manipulation and Adaptive Bilevel Detection," in *Proceedings of Interspeech*, 2025, pp. 5333–5337.

[20] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset," in *Proceedings of the ACM International Conference on Multimedia*, 2024, pp. 7414–7423.

[21] Z. Cai, K. Kuckreja, S. Ghosh, A. Chuchra, M. H. Khan, U. Tariq, T. Gedeon, and A. Dhall, "AV-Deepfake1M++: A Large-Scale Audio-Visual Deepfake Benchmark with Real-World Perturbations," in *Proceedings of the ACM International Conference on Multimedia*, 2025, pp. 13 686–13 691.

[22] Y. Xie, H. Cheng, J. Zhou, X. Guo, T. Wang, J. Liu, W. Wang, R. Fu, X. Wang, H. Huang, X. Huang, L. Ye, and G. Zhai, "AT-ADD: All-Type Audio Deepfake Detection Challenge Evaluation Plan," ACM Multimedia Grand Challenge, 2026. [Online]. Available: https://www.at-add.com/at-add-challenge

[23] H. Delgado, G. Ramondetti, E. Dalmasso, G. Karvitsky, D. Colibro, and H. Talib, "On Deepfake Voice Detection – It's All in the Presentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[24] X. Zhang, Y. Wang, L. Li, L. Jin, and M. Li, "CompSpoof: A Dataset and Joint Learning Framework for Component-Level Audio Anti-Spoofing Countermeasures," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/11465158

[25] X. Zhang, H. Yin, Y. Xiao, L. Zhang, T. Dang, R. K. Das, and M. Li, "Overview of ESDD2: Environment-Aware Speech and Sound Deepfake Detection Challenge," in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2026, accepted.

[26] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A Large-Scale Audio-Visual Dataset," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[27] J. W. Du Bois, W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson, and N. Martey, "Santa Barbara Corpus of Spoken American English, Parts 1–4," Philadelphia: Linguistic Data Consortium, 2005, lDC catalog numbers LDC2000S85, LDC2003S06, LDC2004S10, and LDC2005S25; released 2000–2005. [Online]. Available: https://www.linguistics.ucsb.edu/research/santa-barbara-corpus-spoken-american-english

[28] M. Cartwright, A. Cramer, A. E. Mendez Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, O. Nov, and J. P. Bello, "SONYC-UST-V2: An Urban Sound Tagging Dataset with Spatiotemporal Context," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 75–79. [Online]. Available: https://dcase.community/documents/workshop2020/proceedings/DCASE2020Workshop_Cartwright_68.pdf

[29] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, W. Wang, and M. D. Plumbley, "AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3210–3223, 2024.

[30] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[31] M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon, and B. Shirley, "Preferred Levels for Background Ducking to Produce Esthetically Pleasing Audio for TV with Clear Speech," *Journal of the Audio Engineering Society*, vol. 67, no. 12, pp. 1003–1011, 2019.

[32] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[33] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually Guided Audio Generation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=CYK7RfcOzQ4

[34] V. Negroni, D. Salvi, P. Bestagini, and S. Tubaro, "Analyzing the Impact of Splicing Artifacts in Partially Fake Speech Signals," in *Proceedings of the Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof)*, 2024, pp. 79–85.

[35] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A Database of Multi-channel Environmental Noise Recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013, p. 035081.

[36] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proceedings of Interspeech*, 2019, pp. 1368–1372.

[37] S. Uhlich, G. Fabbro, M. Hirano, S. Takahashi, G. Wichern, J. Le Roux, D. Chakraborty, S. Mohanty, K. Li, Y. Luo, J. Yu, R. Gu, R. Solovyev, A. Stempkovskiy, T. Habruseva, M. Sukhovei, and Y. Mitsufuji, "The Sound Demixing Challenge 2023 – Cinematic Demixing Track," *Transactions of the International Society for Music Information Retrieval*, vol. 7, no. 1, pp. 44–62, 2024.

[38] X. Wang and J. Yamagishi, "A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection," in *Proceedings of Interspeech*, 2021, pp. 4259–4263.