氏名	Lin Zhang	
学位（専攻分野）	博士（情報学）	
学位記番号	総研大甲第2500号	
学位授与の日付	2024年3月22日	
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当	
学位論文題目	“Whether，When，What”：Detection，Localization，and Diarization of Partially Spoofed Audio	
論文審査委員	主査	山岸　順一 情報学コース　教授
		越前　功 情報学コース　教授
		佐藤　いまり 情報学コース　教授
		篠田　浩一 東京工業大学　情報理工学院　教授
		Tomi Kinnunen 東フィンランド大学（University of Eastern Finland） 教授

Summary of Doctoral Thesis

Name in full: Lin Zhang

Title: "Whether, When, What": Detection, Localization, and Diarization of Partially Spoofed Audio

Biometric systems are vulnerable to various manipulations and spoofing, such as text-to-speech synthesis, voice conversion, replay, tampering, adversarial attacks, etc. However, previous research has scarcely explored scenarios where synthetic speech segments are embedded within bona fide speech utterances. We dub this new spoofing scenario as ***"Partial Spoof" (PS)***.

There is an urgent need to explore the PS scenario. It potentially occurs more often, presents higher risks, and is inherently challenging to detect. On the one hand, attackers have various motives to manipulate bona fide speech with segments that are generated through TTS or VC. Such manipulations can drastically alter the intended meaning of an utterance. For instance, by using TTS or VC, an attacker could substitute keywords or phrases, or even introduce negations, effectively changing the original meaning of a sentence. Furthermore, when an attacker has a recording includes partial voice commands, they can generate the missing segments to create complete voice commands and deceive voice-based authentication systems on personal devices. With the advancements in modern TTS and VC technologies, such partial spoofing attacks are becoming more common and present a considerable risk to ASV systems. On the other hand, conventional countermeasures (CMs), often employ aggregation along the whole utterance, bona fide part in the partially spoofed trial would break CM. The urgency to understand and address PS becomes paramount.

In addition, speech anti-spoofing is often viewed as a binary classification task discriminating between spoofed and bona fide speech. However, when it comes to the PS scenario, the pure binary classification is absolutely not enough, as the location and source of the spoofed region in the partially spoofed audio are also important to analyze the intent of the attacker, which requires us to define new tasks, benchmark baselines and metrics beyond the conventional binary classification defined in the traditional speech anti-spoofing field.

Given the significant threat posed by the PS scenario, this thesis undertakes an in-depth exploration of this newly defined PS scenario. It aims to define the PS scenario, construct a database with benchmark models, and analyze the PS scenario based on performance of these benchmark models.

The thesis is organized according to the outline in Figure 1. Chapter 2 describes topics and provides background related to this thesis. Chapter 3 defines the under-explored Partial Spoof with motivation and constructs benchmark, including a database with detailed annotation, three tasks for PS, and their corresponding metrics. Then, Chapter 4 to 8 propose potential solutions aimed at constructing the benchmarks to explore and analyze the afore-mentioned three tasks for the PS scenario. Finally, Chapter 9 concludes this thesis and highlights potential directions for future research and improvements.

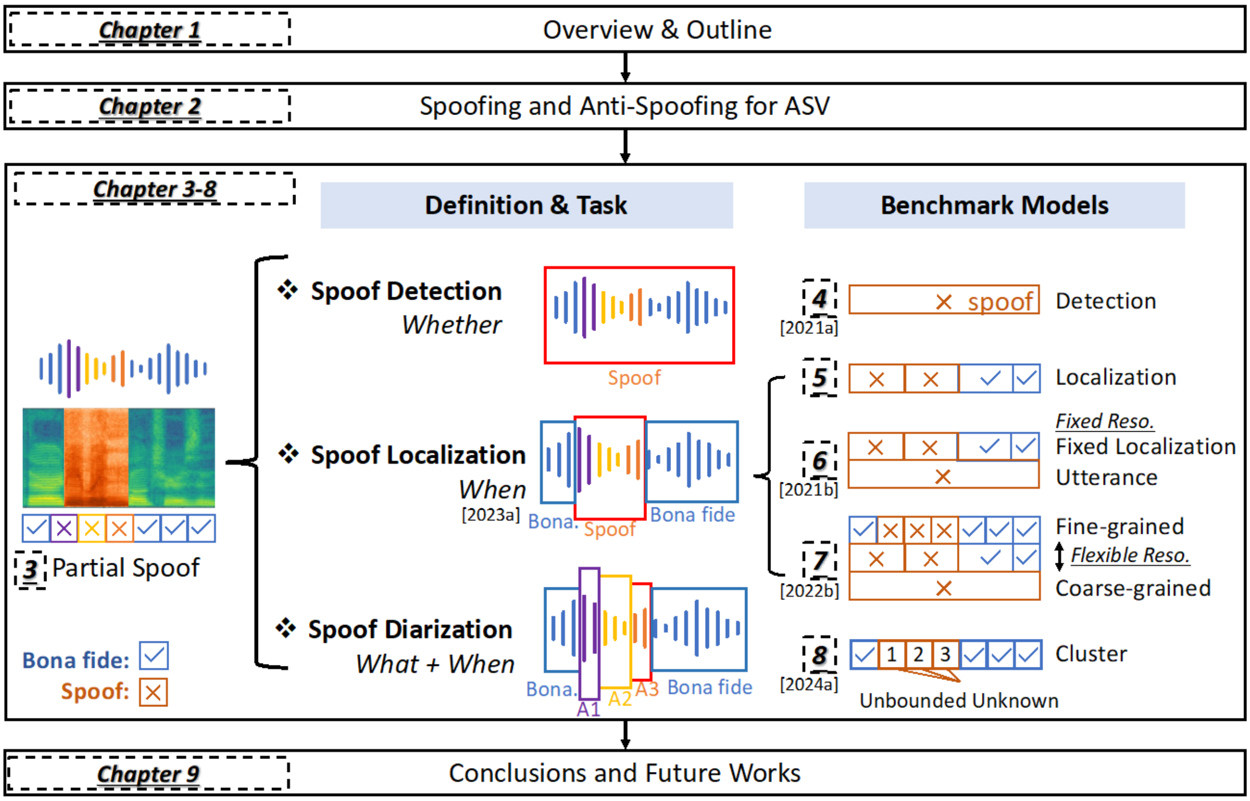

Figure 1. Thesis outline. Reference in '[]' represents the author's publications, which can be found in Appendix.

As one of the pioneering studies on the PS scenario, we designed three specific tasks, each associated with a question, in the Chapter 3:

- **Spoof Detection:** Whether the utterance is spoofed?
  This task aligns with the common task in the spoofing community distinguishing whether an utterance is spoofed or bona fide.
- **Spoof Localization:** When do spoofs happen?
  This task aims to determine the location of spoof segments within utterances.
- **Spoof Diarization:** What attacks when?
  This task not only locates the spoofed segments but also discriminates the specific spoofing techniques employed.

From chapter 4 to 8, to explore the afore-mentioned tasks presented by the PS scenario, we developed various CMs, from conventional models enhanced with advanced strategies to cutting-edge models using self-supervised learning (SSL) for training:

1. A light convolutional neural network (LCNN) as single-task-learning baseline for spoof detection (Chapter 4) and localization (Chapter 5).
2. Squeeze-and-excitation LCNN (SELCNN)-based fixed-resolution (Chapter 6) and SSL-based flexible-resolution (Chapter 7) models with multi-task learning for simultaneous spoof detection and localization.
3. A CM-Conditional Clustering model for spoof diarization (Chapter 8).

Besides, this thesis derives several key conclusions with the potential to provide valuable insights for future research in the PS scenario:

1. Spoof detection on the PS scenario is more challenging than on the fully spoof scenario (Chapter 4), and spoof localization is more complex than detection but is feasible (Chapter 5). Spoof diarization is an emerging and complex task, we managed to achieve promising results under simplified conditions (Chapter 8).
2. Using PS training data is beneficial. CMs trained using partially spoofed audio exhibit robustness against both PS and fully spoof. However, CMs trained solely on fully spoofed data would easily fail when facing the PS scenario. (Chapter 4)
3. Spoof detection and localization are closely related to each other, and they can perform each other's tasks without additional labeling or training. (Chapter 5)
4. Utterance-level spoof detection can benefit from the use of more fine-grained information, underscoring the importance of precise timestamp annotations for training. For spoof localization, training models to focus on finer temporal resolutions yield better performance. (Chapter 6 and 7)
5. CMs trained with different bona fide and spoof labeling schemes capture distinct information. Properly integrating different CMs can improve performance on the spoof diarization (Chapter 8).

In summary, this thesis establishes a series of benchmarks for the PS scenario research, representing a remarkable contribution to the speech anti-spoofing community. It serves as a foundation for further investigation into the PS scenario and is the first study to release comprehensive PS-related resources. This includes a database with detailed timestamp annotations on spoofing methods, codes, and models. All related resources, as mentioned in this PhD thesis are available at https://github.com/nii-yamagishilab/PartialSpoof .

（別紙１ Separate Sheet 1）

Results of the doctoral thesis defense

博士論文審査結果

氏　　名（Name in Full）　Lin Zhang

論文題目（Title）　“ Whether, When, What”: Detection, Localization, and Diarization of Partially Spoofed Audio

本学位論文は、「“Whether, When, What”: Detection, Localization, and Diarization of Partially Spoofed Audio」と題し、全９章で構成されている。本論文は、合成音声の一部が自然音声に挿入され、なりすましやディープフェイクとして悪用されるPartialSpoofを新たな脅威対象として捉え、その様な音声に対してReal/Fake判定を行う検出技術、合成音声が挿入された時間領域を特定するローカライゼーション技術等を提案し、研究成果をまとめたものである。

第１章では本論文で扱う問題の重要性、位置付けおよび貢献について説明がなされ、第２章では話者認識システムの概要、なりすましおよびその対策技術の概要、およびそれらの指標について説明がなされた。第３章では、PartialSpoofのシナリオを定義し、検出タスク、ローカライゼーションタスク、ダイアリゼーションタスクという３つのタスクの説明がなされた。また学習用データベースの構築ポリシーや手順についての詳細が示された。次に、第４章では、発話単位で検知を行うモデルを提案し、実験からその有効性を示した。第５章では合成音声が挿入された時間領域を特定するローカライゼーション用のモデルを提案し、実験からその有効性を示した。第６章では発話単位で検知を行うモデルとローカライゼーション用のモデルとを同時に学習するマルチタスク学習の提案がなされ、第７章ではさらに複数の異なる時間解像度において検知モデルとローカライゼーションモデルとをマルチタスク学習する手法の提案があり、実験からその有効性を示した。第８章ではさらに、検出されたセグメントをグルーピングするダイアリゼーションタスクのモデルの提案と実験結果の報告もなされた。９章ではまとめと今後の課題を述べている。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会および口述試験では、審査員からの質疑に対して適切に回答がなされた。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では出願者の博士研究に学位論文として十分なレベルの新規性や有効性があると評価した。

以上を要するに、本学位論文は、音声セキュリティーおよびメディアフォレンジック分野を学術的に発展させる内容であると同時に、音声情報処理を利用したサービス等にも直結する内容であり、その科学的貢献は大きいと言える。また本学位論文の成果は学術雑誌論文１編、国際会議論文３編として発表され、社会的な評価も得ている。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。