氏　　名	Chang Zeng
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2499 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	Spoofing-aware Speaker Verification System Robust Against Domain and Channel Mismatches
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 小川　哲司 早稲田大学　理工学術院　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: Chang Zeng

Title: Spoofing-aware Speaker Verification System Robust Against Domain and Channel Mismatches

Automatic speaker verification (ASV) has shown immense potential across domains like security, forensic analysis, and human-computer interaction. However, real-world deployment necessitates ASV systems that are robust to diverse conditions involving channel variability, spoofing attacks, and domain mismatch between training and test environments. This thesis has focused on a pivotal research problem - enhancing the robustness of automatic speaker verification (ASV) systems to enable reliable deployment in the real world. Through a combination of novel methods for addressing channel mismatch, spoofing attacks, and domain mismatch, it has made significant contributions towards this challenging goal.

The primary objectives of this thesis have been:

1. To develop techniques that improve the robustness of ASV systems to channel mismatch between enrollment and test conditions.
2. To enhance resilience against spoofing attacks like synthesized and converted speech.
3. To tackle the universal issue of domain mismatch between training and deployment environments.
4. To address these three threats jointly in an integrated manner.

Through a combination of the pair-wise learning paradigm, spoofing attack simulation, and meta-learning paradigm, the author has made notable advancements in each of these research goals. The highlights of contributions are summarized below.

To mitigate the impact of channel variability on ASV reliability, the author proposed an attention back-end model in Chapter 3 that directly handles multiple enrollment utterances. Unlike conventional averaging of embeddings, this architecture employs two self-attention networks to explicitly model relationships among enrollment utterances - scaled dot self-attention and feed-forward self-attention. The pair-wise learning paradigm trains the model by sampling varied combinations of conditions from the dataset, realistically simulating channel mismatch.

Experiments on VoxCeleb and CNCeleb datasets demonstrate that the attention back-end consistently achieves lower EER and minDCF than the widely used PLDA back-end. On CNCeleb, it reduces EER from 12.52% to 10.12% and minDCF from 0.6105 to 0.5649. More significantly, in the genre mismatch scenario, the attention backend displays greater robustness than PLDA for most cross-genre test cases.

Through explicit modeling of enrollment utterances and realistic simulation of variability, the attention back-end significantly enhances ASV's reliability against channel mismatch.

While channel robustness is critical, ASV systems face another key threat - synthesized and converted speech attacks. Chapter 4 tackles this challenge by extending the attention back-end to become spoofing-aware.

The core innovations lie in fusing speaker verification and countermeasure modules through score-level integration, along with refining the pair-wise learning strategy to incorporate spoofing attacks during training. By allowing gradients to flow between the scores, interactions between the modules are induced, overcoming the limitations of isolated training. The composition of training trials is also enlarged to encompass bonafide, spoofed, and non-target examples.

Evaluated on the SASV Challenge 2022 benchmark, the spoofing-aware model slashes the SASV-EER from 22.91% to 1.19%, showcasing tremendous improvement in resilience. The SV-EER remains competitive at 1.32%, confirming that ASV accuracy in normal conditions is uncompromised. Detailed ablative tests highlight the significance of each component in achieving spoofing robustness.

The problem of distribution differences between training and test sets affects all machine learning systems, including ASV. The author tackles this in Chapter 5 using a meta-learning paradigm. The key intuition is to simulate domain mismatch during training by constructing distinct meta-train and meta-test splits with genre variability. Additional genre alignment regularization is imposed through adversarial training.

Experiments on a new cross-genre anti-spoofing task validate the approach, with EER reductions of up to 26.7% on unmatched genres and consistent gains over multiple protocols. For instance, on CGP-I with the “singing” genre unseen during training, meta-learning improves EER from 9.517% to 8.248%. The addition of genre alignment further boosts it to 8.157%. This demonstrates the efficacy of the proposed method for domain generalization.

While Chapter 3 to Chapter 5 tackles the three challenges independently, chapter 6 proposes an integrated framework combining all techniques. Pair-wise learning and spoofing simulation handle channel and attack issues, while meta-learning provides domain robustness. This unified approach is engineered to handle diverse real-world conditions.

The integrated results are highly promising, with SV-EERs and SASV-EERs considerably lower than the baseline systems with the simple supervised learning paradigm. For example, on the CNComplex testing dataset, the integrated learning paradigm achieves a SASV-EER of 7.37% on the cross-genre protocol I, compared to 37.64% SASV-EER of ECAPA-TDNN model with the supervised learning paradigm.

In summary, through a diverse set of innovations, this thesis substantially advances the state-of-the-art in robust ASV under challenging conditions. The consistent

performance gains provide empirical evidence of the effectiveness of our methodology in tackling key bottlenecks toward reliable ASV deployment. By enhancing real-world robustness, this thesis marks an important milestone in progressing voice biometrics from the lab towards ubiquitous adoption across applications.

（別紙１ Separate Sheet 1）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏　　名　Chang Zeng

Title
論文題目　Spoofing-aware Speaker Verification System Robust Against Domain and Channel Mismatches

本学位論文は、「Spoofing-aware Speaker Verification System Robust Against Domain and Channel Mismatches」と題し、全７章で構成されている。本論文は、音声により個人認証を行う話者認識技術において、エンロールデータと呼ばれるユーザ登録データと実際のテストデータのミスマッチの影響、学習データとテストデータのミスマッチの影響、なりすまし攻撃の影響を同時に考慮し、これら全てに対してモデルを頑健にする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､および､話者認識システムにおけるエンロールデータ、学習データ、テストデータ間のミスマッチにより起こる問題とその対処法について説明がなされた。第３章では、まずエンロールデータとテストデータのミスマッチの影響を軽減するモデルの構造が提案され、その有効性が実験から示された。第４章では、第３章で提案したモデルをさらに拡張し、なりすまし攻撃の影響も同時に軽減する枠組みを提案し、その有効性を示した。次に、第５章では学習データとテストデータのミスマッチの影響を軽減するメタ学習の枠組みを新たに提案し、その有効性を示した。第６章では３章～５章で提案された個別の手法を統合することで、エンロールデータとテストデータのミスマッチの影響、学習データとテストデータのミスマッチの影響、なりすまし攻撃の影響を全て同時に考慮するシステムとそのモデル構造を提案し、実験からその有効性を示した。７章ではまとめと今後の課題を述べた。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、話者認識および音声セキュリティー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。