中国科学技术大学

博士学位论文

论文题名: Robust Speaker Verification Under Channel Mismatch

学生姓名: Wenjie Cao

导　　师: Prof. Han Zhao

提交日期: 二〇二六年八月二十七日

中图分类号：____ 学校代码：10358
UDC 分类号：____ 密　　级：____

中国科学技术大学

博士学位论文

(学术学位)

Robust Speaker Verification Under Channel Mismatch TitleoftheThesis

学生姓名：Wenjie Cao
学　　号：SA26010001
导师姓名：Prof. Han Zhao
院　　系：School of Information Science and Technology
学科专业：Signal and Information Processing
研究方向：Speech and Audio Processing
论文提交日期：二〇二六年八月二十七日

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：_____ 签字日期：_____

中国科学技术大学学位论文使用授权声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致，如因不符而引起的学术声誉损失由本人自负。

控阅的学位论文在解除控阅后也遵守此规定。

☑公开 ☐控阅（____ 年）

作者签名：_____ 导师签名：_____

签字日期：_____ 签字日期：_____

研究生学位论文（实践成果报告）未使用人工智能工具声明

本学位论文（实践成果报告）的研究与撰写过程未使用任何生成式人工智能工具（指能够根据提示自动生成文本、代码、图表等内容的大模型及其衍生工具）。

所有内容均基于个人研究、文献阅读及实验数据完成，本人承诺符合学术诚信要求，论文中不存在任何人工智能辅助生成内容。违反本声明所产生的责任与后果，由本人独立承担。

作者签名：_____

日期：____年__月__日

摘要

Speaker verification systems trained on clean, matched-condition audio often degrade sharply when deployed on recordings from a different microphone or transmission channel than the one used during training. This thesis studies this channel mismatch problem and proposes a representation learning method that explicitly separates speaker identity from channel characteristics during training, using a small amount of multi-channel data recorded from the same speakers. We evaluate the method on a benchmark constructed from recordings across five distinct channel conditions and show that it substantially reduces the equal error rate relative to a standard x-vector baseline under channel mismatch, while matching baseline performance in the matched-channel condition.

关键词：

Contents

Chapter 1 Introduction 1
1.1 Motivation 1
1.2 Contributions 1
1.3 Thesis Outline 1
Chapter 2 Related Work 2
2.1 Deep Speaker Embeddings 2
2.2 Channel and Domain Robustness 2
2.3 Gap Addressed by This Thesis 2
Bibliography 3
Acknowledgments 4

List of Figures

List of Tables

Chapter 1 Introduction

Speaker verification – determining whether two recordings were spoken by the same person – has become a standard component of voice-based authentication systems. Modern systems based on deep speaker embeddings, such as x-vectors, achieve low error rates when training and test audio come from matched recording conditions [1].

1.1 Motivation

In practice, enrollment and verification audio are frequently captured on different devices or transmitted over different channels – a phone enrolled over a landline, verified later over a mobile network, for example. This channel mismatch introduces a systematic shift in the learned speaker embedding space that is difficult to distinguish from a genuine speaker difference [2].

1.2 Contributions

This thesis makes the following contributions:

- A representation learning method that separates speaker identity from channel characteristics using paired multi-channel recordings of the same speakers.
- A benchmark constructed from recordings across five distinct channel conditions, used to evaluate verification performance under controlled channel mismatch.
- An evaluation showing substantially reduced equal error rate under mismatch relative to a standard x-vector baseline, with no loss of accuracy in the matched condition.

1.3 Thesis Outline

2 reviews prior approaches to channel-robust speaker verification. Later chapters (omitted from this starter document) would present the proposed method, the benchmark construction, and experimental results in detail.

Chapter 2 Related Work

2.1 Deep Speaker Embeddings

X-vector embeddings, trained with a classification objective over speaker identity and pooled across an utterance, remain a strong baseline for speaker verification[1].

2.2 Channel and Domain Robustness

Joint factor analysis and related generative approaches explicitly model channel variability as a separate latent factor from speaker identity, predating the deep embedding era but sharing the same underlying goal[2].

2.3 Gap Addressed by This Thesis

Most channel-robustness work for deep embeddings relies on large amounts of unpaired multi-channel data and domain adversarial training. This thesis explores a paired-data setting, where a modest number of speakers have recordings across multiple channels, which allows more direct supervision of what should and should not change across channels.

Bibliography

[1] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: robust DNN embeddings for speaker recognition[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2018: 5329-5333.

[2] Kenny P. Bayesian speaker verification with heavy-tailed priors[J]. Odyssey Speaker and Language Recognition Workshop, 2010.

Acknowledgments

I would like to thank my supervisor for the guidance and support offered throughout this project, and my colleagues in the Speech and Audio Processing group for many helpful discussions and for their assistance with data collection.