厦门大学

硕士学位论文

基于 PLDA 模型的说话人识别方法研究

Research of Speaker Recognition Based On PLDA Model

黄玲

指导教师姓名：李琳 副教授

专业名称：电路与系统

论文提交日期：2015 年 4 月

论文答辩时间：2015 年 5 月

学位授予日期：2015 年 月

答辩委员会主席：

评阅人：

2015 年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：黄珍

2015 年 5 月 24 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：黄玲

2015 年 5 月 24 日

摘　要

说话人识别技术，作为现代重要的生物信息识别技术之一，通过对说话人语音样本提取的特征参数进行建模分类，从而分辨说话人身份。目前，NIST(National Institute of Standards and Technology)国际评测结果显示，基于 PLDA(Probabilistic Linear Discriminant Analysis)模型的说话人识别系统可获得突出的识别效果。然而，现实生活中，语音样本很容易受到环境噪声的干扰，有时候注册语音和待测语音的样本时长是不一致的，甚至，在某些信道较难采集到丰富的语音样本数据以供 PLDA 模型训练，上述这些复杂问题，在一定程度上制约了基于 PLDA 说话人识别系统的实际应用。因此，对基于 PLDA 模型的说话人识别技术进行研究是非常有意义的。

本文主要针对说话人识别系统中语音增强、时长不匹配和训练样本有限这三个问题，分别提出了有效的解决方案。本文的主要工作及创新点如下：

1、基于稀疏表示原理，提出了自适应更新字典的双稀疏语音增强方法，并且，使用与待处理带噪语音无关的干净语音经 K-SVD(K-Singular Value Decomposition)训练统一字典原子，在确保语音增强效果的前提下提高了计算效率；

2、引入语音时长调整 PLDA 模型分布的协方差参数，提出了基于时长约束的概率修正 PLDA 的说话人识别方法，提高了说话人识别系统在时长不匹配时的识别效果；

3、利用大样本信道训练得到的 PLDA 参数为先验值，提出 PLDA 参数更新的跨域迁移策略，以解决小样本信道 PLDA 建模的困难，并在语音样本有限的前提下提高了识别性能。

本论文的研究工作还包括一个跨信道说话人识别语音库的建设。该语音库采集了 100 个说话人语音样本，含有丰富的信道信息（耳麦、会议麦、录音笔、固话信道、两路移动电话信道、网络语音信道等），即，对每个说话人同步在八个信道录制相同文本内容。这个数据库的建立对进一步研究跨信道说话人识别及文本相关说话人识别系统都具有重要意义。

关键词：说话人识别；PLDA；时长不匹配；跨域迁移

Abstract

Speaker recognition as an important technology in modern biological information recognition area, it recognizes the speaker through analysing and modeling speaker's voice characteristics. Currently, NIST (National Institute of Standards and Technology) international evaluation results show that speaker recognition system based on PLDA model achieves outstanding recognition results. However, in real-world conditions, speech is susceptible to environmental noise, and sometimes the duration of the target speech and test speech is mismatch, even, it is difficult to collect enough data for PLDA(Probabilistic Linear Discriminant Analysis) training in some channels. Due to the above-mentioned problems, the performance of speaker recognition system based on PLDA model decline dramatically.

This paper mainly focused on the speech enhancement, duration mismatch and limited training samples issues in speaker recognition, and proposed effective solutions respectively. The main work and innovation of this paper is as follows:

1. Based on sparse representation theory, a double sparsity speech enhancement method was proposed, which can update the dictionary adaptively. A practical method using a universal dictionary, which was trained on many irrelevant clean speech utterances by K-SVD (K-Singular Value Decomposition) algorithm, was proposed to improve the enhancement performance and accelerated the computational speed.

2. A modified-prior PLDA model was proposed to deal with the duration mismatch issue in speaker recognition system. By revising the covariance parameters with speech duration information to adjust the distribution of the PLDA model, the robust performance of speaker recognition system was obtained under the duration mismatch condition.

3. An cross-domain transfer strategy was proposed, in which PLDA parameters of in-domain speech samples can be updated adaptively by MAP method on the prior knowledge of the PLDA parameters of out-domain speech samples. As a result, the difficulties for small samples to train the PLDA model would be resolved.

Research work of this paper also includes the construction of a cross-channel speaker recognition database. This database collected 100 speakers' speech samples, including rich channel information (headsets, conference wheat, voice recorder, fixed channel, two kinds of the mobile telephone channels, network voice channel). The

same content was recorded simultaneously in eight channels for each speaker. Further study on the cross-channel speaker recognition and text-dependent speaker recognition would be encouraged on this inter-session database.

Keywords: speaker verification; PLDA; duration mismatch; cross-domain transfer

目录

CONTENTS

第一章 绪论

1.1 说话人识别概述

1.1.1 说话人识别的研究背景与意义

随着社会经济的发展，网络信息化和智能化技术和应用的普及，身份认证问题也就变得非常重要。传统的身份认证方法有通过密码、身份证、条形码等物品来认证身份，这类方法一般是利用能标识个人身份的物品或口令来认证身份，存在容易被非法用户盗取、破解，个人携带物品容易遗失等问题，传统的身份认证方法并不是那么的安全可靠，于是，人们开始探索使用生物识别技术进行身份认证。生物识别技术一般是使用人体的某个或多个生物特征进行身份认证的，由于人体的生物特征具有唯一性和不可复制性，这一生物密钥不易被盗取、破解或遗忘，因此利用生物识别技术来进行身份认证比传统的身份认证方法更有安全性、保密性和便捷性。说话人识别作为一种生物识别技术，和虹膜识别、指纹识别、掌纹识别等一样，有无须记忆和使用便捷等优势，它能够满足人们对于身份识别的安全性、实用性、准确性的要求。

和其他生物识别技术相比较，说话人识别技术表现出很多独特的优势和优良的特性：

首先，每个说话人的声音区分性非常强。即使是双胞胎，他们的声音特征也不可能是完全一样的。声音的独特性和唯一性使得把声音作为生物特征进行身份识别成为可能。

其次，说话人识别技术方便快捷。声音的获取自然、方便，声音提取可在不知不觉中完成，把声音作为识别特征，用户容易接受，不涉及用户隐私。在说话人识别技术中，用户不用像指纹识别那样将手指放在传感器上，也不必像人脸识别那样把眼睛对着摄像头，只要简单地说一两句话就可以完成识别。

第三，说话人识别技术中所使用的设备成本比较低，用户容易使用，并且还可以远程操作。通常使用简单的麦克风、电话、手机等设备就可以

完成声音的采集；声音的采样与量化也比较简单，对芯片也没有特殊要求；特征提取和模型的训练和匹配都只要在普通的计算机上就可以完成。

由于上述优点，说话人识别技术在日常生活的很多领域都有着重要的应用，可用于各种需要进行身份识别的安全领域：

(1) 门禁系统。例如家庭、机场通道和特殊安全入口等，都可以使用通过声音进行身份识别的声纹锁。

(2) 金融系统。随着远程炒股、电话银行等业务的不断增加，用户只需要使用语音密码，就可以安全、有效地实现身份的确认。

(3) 司法鉴定。在许多民事、刑事诉讼和案件的侦查过程中，通常需要通过说话人识别技术对录音、电话通话等证据的做分析，利用说话人识别技术来协助案件的审理。

(4) 远程身份认证：如在如声音拨号、远程数据库访问和计算机远程登入等领域，通常可以使用说话人识别技术实现远程身份认证。

1.1.2 说话人识别的发展与现状

说话人识别技术的研究最早始于 20 世纪 30 年代，当时的研究主要集中于人耳听辨实验和探讨听声音识别的可能性方面。随着“声纹(Voiceprint)”概念的提出，说话人识别技术的研究开始进入数字化处理与分析的时代。在 60 年代之后，随着理论研究和计算机科学技术的深入发展，说话人识别技术研究工作向智能化、自动化方向发展。此时学者们主要关注于对特征的提取和改进，并把倒谱分析和线性预测分析[1]等方法使用到说话人识别中。70 年代后，说话人识别技术的研究重点开始向说话人个性特征的分离与增强、对于声学特征参数处理和新的模式匹配方法[2]等方面发展。至 80 年代后，隐马尔可夫模型[3]和人工神经网络[4]在语音识别技术领域的成功应用，促使学者们将这些技术引进到说话人识别领域中，在文本相关的说话人识别领域中获得了很好的识别效果。到 90 年代后，Reynolds 对先后提出了高斯混合模型(Gaussian Mixed Model, GMM)[5,6]，和通用背景模型(Universal, Background Mode, UBM)[7]，从说话人语音特征分布的角度出发，使用一些非监督学习的方法对说话人语音特征分布进行拟合，这种方法提高了说话人识别中类别区分能力，迅速成为说话人识别领域中的主

流技术。同时，学者们利用支持向量机(Support Vector Machine，SVM)[8,9,10,11]的显著区分机制，使用不同的核函数实现了不少优秀的说话人识别系统。

近年来，涌现了不少新的说话人识别技术，如 SVM 与 GMM 结合[12]、多模态识别[13]、语音高层信息的探讨[14]、以及针对信道失配问题的说话人模型合成技术(Speaker Model Synthesis, SMS)[15]等。到 2008 年以后，超矢量(Supervector)技术成为说话人识别技术的一个新的研究热点和方向。通常我们是用一个特征矢量集来表征一段语音，该集合的维度受语音时长的影响而变化，而超矢量技术是使用一个固定大小的高维单向量(超矢量)来表征语音。在这个基础上，Kenny 提出了联合因子分析(Joint Factor Analysis, JFA)技术[16,17]。JFA 技术把语音的差异性分为两个子空间：说话人与说话人之间的差异 (Speaker Variability)和相同说话人不同段语音的差异(Session Variability/Channel Variability)。目前，JFA 已经发展成为说话人识别的主流技术之一。受 JFA 的启迪，Dehak 提出了 i-vector (Identity Vector)思想[18]，把不同语音间的差异性用一个更低维的子空间表示，和超矢量相比，i-vector 是一个用来表示各种长短的语音段更低维的向量。通常，基于 i-vector 的说话人识别系统需要和一些信道补偿技术，例如 WCCN(Within-class Covariance Normalisation)[19]和 LDA(Linear Discriminant Analysis)[20]结合使用。接着，Kenny 又提出将概率线性鉴别分析(Probabilistic Linear Discriminant Analysis, PLDA)模型应用在说话人识别中[21]。PLDA 模型首先是由 Prince 提出并应用于人脸识别[22]，它是传统 LDA (Linear Discriminant Analysis)[23]技术的概率形式。当前，PLDA 模型在注册和测试信道不匹配时的说话人识别中表现出很强的鲁棒性。

1.2 说话人识别技术面临的问题

当前，实验室环境下(干净语音)的说话人识别技术已经发展得很成熟，然而现实环境中的说话人识别技术仍然面临很多复杂的问题，主要包括以下几个方面：

1)说话人是声音会随着环境、情绪、健康等状况的变化发生变化的。而且语音还具有长时变动特性，和时间及年龄有关；

2)声音可以被模仿、合成、转录，形成“冒充”身份攻击说话人识别系

统；

3)声音通过通讯设备传输时，会受到传输设备所产生的噪声干扰，不同的通讯设备噪声情况可能不同的。同时，环境噪声也是无处不在的，它们严重影响了系统的识别率；

4)时长不匹配是说话人识别技术实用化过程中必须解决的问题。在身份识别系统中，用户的注册语音与测试语音往往是不等时长的，一般情况下，注册语音会比较长，而测试时，特别是在信息侦查应用中，能够获取的语音长度往往很有限；

5)跨信道问题也是影响说话人识别技术的关键问题之一，实际语音的信道环境是复杂多样的，很难保证采集到同一信道的样本数据都是丰富的，对小样本数据库的说话人识别难度较大。

本文侧重针对说话人识别技术中的语音增强、时长不匹配和小样本数据等问题展开研究工作，分别提出有效的解决方案。

1.3 论文主要内容和结构

为了推进说话人识别技术的实用化，本文主要针对语音增强、时长不匹配和小样本数据库情况下的说话人识别技术及相关理论和算法做了进一步的研究和探索。本文分析比较了多种主流语音增强方法，提出了一种基于稀疏表示的语音增强方法，具有较好的降噪效果。针对注册语音与测试语音时长不一致情况，本文提出一个新的联合时长信息的概率修正 PLDA 建模方法，提高 PLDA 对每个说话人每段语音的时长表征能力，以增强说话人类别的区分度。本文建立了一个八信道说话人数据库，并在此数据库上探索了 PLDA 参数更新的跨域迁移策略，以论证小样本数据库识别性能提升的可行性。

论文结构的具体安排如下：

第一章是论文的绪论部分，主要阐述了说话人识别技术的研究背景和研究意义，以及国内外说话人识别技术的研究现状和存在的问题，最后介绍了本论文的主要工作和论文的组织结构。

第二章阐述了说话人识别技术的理论基础和关键技术。首先介绍了说话人识别的基本原理，然后详细描述了语音信号的预处理，特征提取，说