

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): 

2015年 4月 20日



Y2926738

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）：王唯

2015年5月20日

摘要

说话人识别技术作为生物特征识别的一个方向，在理论和应用领域都得到了快速发展和广泛关注。本文主要是对基于 i-vector 的说话人识别算法进行研究，在此基础上引入了 PLDA 信道补偿算法以及海量数据的无标注聚类。

在对经典说话人识别算法的介绍中，从理论和实际应用中分析了经典算法存在的缺陷，由此推出了基于 i-vector 的说话人识别算法，阐述了算法的原理和工作流程，并通过实验证明了其优越的性能。为了提高系统的鲁棒性，本文研究了 PLDA 信道算法，对算法的原理进行了分析，从理论上说明了 PLDA 算法对信道补偿的作用。

在实际应用中，存在着大量的无标注数据，而与此同时 PLDA 等算法需要大量的有标注数据进行训练。基于这样的矛盾，本文重点研究了在无标注数据下的说话人识别技术。以 NIST2014 说话人评测为出发点，首先介绍了无标注数据用于 SVM 分类系统中对系统分类性能的提升，然后阐述了基于 AHC 的三种合并策略的聚类算法，并利用 AHC-PLDA 聚类算法为无标注数据进行标记，最后将标记后的数据运用到 PLDA 参数训练中，通过实验证明了基于 AHC-PLDA 聚类和 PLDA 组合系统对识别性能提升明显。

最后，结合实际的语音排查项目，分析了基于 i-vector 说话人识别技术在应用中的优势，实现了基于说话人识别的语音排查系统。

关键词：说话人识别；i-vector；PLDA；无监督聚类

Abstract

Speaker recognition, as one important technology of biometric authentication, has gained great progress in both theory and applications. This study focuses on the state-of-the-art i-vector system. And the channel compensation algorithm and unsupervised clustering technology of large-scale data are also studied.

We first give the basic introductions of classic speaker recognition algorithms, and analyze their defects. After that, we have a deep study of i - vector, including its principle and working process. In order to improve the robustness of inter-session, we also study the probabilistic linear discriminant analysis (PLDA) algorithm, which has quite good performance of channel compensation.

In practice, there are large-scale unlabeled data, which can be used directly by PLDA. To utilize large-scale data, we have further study of unsupervised clustering. Based on NIST2014 speaker evaluation data, we apply support vector machine (SVM) and agglomerative hierarchical clustering (AHC) clustering algorithm respectively. The labeled data after clustering are used for PLDA training. Experimental results show that the combined system based on AHC-PLDA clustering and PLDA verification improved the performance greatly, compared with the baseline Cosine system.

Finally, we design and realize an i-vector system for the application of speech analysis, which requires finding the similar speakers in large-scale speaker data. Detailed progress of this analysis system is described.

Keywords: Speaker Recognition; i-vector; PLDA; Unsupervised Clustering

目 录

| | |
|---|------------|
| 摘要..... | I |
| Abstract..... | III |
| 第一章 绪论..... | 1 |
| 1.1 研究背景及意义..... | 1 |
| 1.2 说话人识别技术研究及现状..... | 2 |
| 1.2.1 说话人识别技术介绍..... | 3 |
| 1.2.2 说话人识别技术研究现状..... | 5 |
| 1.3 本文主要工作..... | 7 |
| 1.4 本文结构安排..... | 7 |
| 第二章 说话人识别系统概述..... | 9 |
| 2.1 引言..... | 9 |
| 2.2 说话人识别系统的基本结构..... | 9 |
| 2.2.1 有效语音检测..... | 10 |
| 2.2.2 特征提取..... | 11 |
| 2.3 GMM-UBM 说话人识别系统 | 14 |
| 2.3.1 GMM 简介..... | 14 |
| 2.3.2 GMM-UBM 的训练..... | 16 |
| 2.3.3 GMM-UBM 似然比得分 | 17 |
| 2.4 GMM-SVM 说话人识别系统..... | 17 |
| 2.4.1 SVM 简介 | 18 |
| 2.4.2 GMM-SVM 说话人识别系统框架 | 18 |
| 2.4.3 SVM 系统在说话人识别应用中面临的问题 | 20 |
| 2.5 本章小结..... | 21 |
| 第三章 基于 i-vector 的说话人识别系统 | 23 |
| 3.1 引言 | 23 |
| 3.2 基本思想 | 23 |

| | |
|--|-----------|
| 3.3 基于 i-vector 模型的说话人识别系统..... | 24 |
| 3.3.1 全局差异空间的估计 | 25 |
| 3.3.2 i-vector 的估计 | 26 |
| 3.3.3 系统测试打分 | 27 |
| 3.3.4 系统性能分析评估 | 28 |
| 3.4 信道补偿的 PLDA 算法..... | 30 |
| 3.4.1 PLDA 模型参数的训练 | 31 |
| 3.4.2 基于 PLDA 模型的确认得分 | 33 |
| 3.4.3 PLDA 参数的 MAP 自适应..... | 34 |
| 3.4.4 实验结果与分析 | 35 |
| 3.5 本章小结..... | 36 |
| 第四章 无监督聚类和说话人识别 | 39 |
| 4.1 引言 | 39 |
| 4.2 NIST2014 评测 | 39 |
| 4.2.1 NIST2014 数据介绍 | 40 |
| 4.2.2 Whitening 规整 | 41 |
| 4.3 无标注数据的聚类实验与分析..... | 42 |
| 4.3.1 实验数据 | 42 |
| 4.3.2 基于 SVM 算法子系统 | 43 |
| 4.3.3 AHC 聚类算法..... | 45 |
| 4.3.4 AHC+PLDA 子系统 | 50 |
| 4.4 系统得分融合 | 54 |
| 4.5 本章小结..... | 56 |
| 第五章 说话人识别应用实例..... | 59 |
| 5.1 项目意义 | 59 |
| 5.2 项目方案 | 59 |
| 5.3 小结 | 65 |
| 第六章 全文总结及工作展望..... | 67 |

| | |
|------------------|----|
| 参 考 文 献..... | 69 |
| 攻读硕士期间的科研成果..... | 73 |
| 致 谢 | 75 |

Table of Contents

| | |
|---|------------|
| Abstract in Chinese | I |
| Abstract in English..... | III |
| Chapter 1 Introduction | 1 |
| 1.1 Research Background | 1 |
| 1.2 Current Issues of Speaker Recognition | 2 |
| 1.2.1 Summary of Speaker Recognition..... | 3 |
| 1.2.2 Research Progress of Speaker Recognition..... | 5 |
| 1.3 Main Works of This Study..... | 7 |
| 1.4 Structure of This Study..... | 7 |
| Chapter 2 Overview of Speaker Recognition System | 9 |
| 2.1 Introduction | 9 |
| 2.2 Basic Structure of Speaker Recognition System | 9 |
| 2.2.1 Voice Active Detection..... | 10 |
| 2.2.2 Feature Extraction | 11 |
| 2.3 Speaker Recognition System Based on GMM-UBM | 14 |
| 2.3.1 GMM..... | 14 |
| 2.3.2 The Training Process of GMM-UBM | 16 |
| 2.3.3 Likelihood of GMM-UBM..... | 17 |
| 2.4 Speaker Recognition System Based on GMM-SVM..... | 17 |
| 2.4.1 SVM | 18 |
| 2.4.2 GMM-SVM Framework | 18 |
| 2.4.3 Issues of SVM System | 20 |
| 2.5 Summary | 21 |
| Chapter 3 Speaker Recognition System Based on I-vector | 23 |
| 3.1 Introduction | 23 |
| 3.2 Basic Thought | 23 |
| 3.3 Speaker Recognition System Based on I-vector Model | 24 |
| 3.3.1 Total Variability Space | 25 |
| 3.3.2 i-vector Evaluation | 26 |
| 3.3.3 Score Evaluation | 27 |

| | |
|--|-----------|
| 3.3.4 Performance of System | 28 |
| 3.4 PLDA Algorithm..... | 30 |
| 3.4.1 PLDA Training | 31 |
| 3.4.2 PLDA Score..... | 33 |
| 3.4.3 PLDA MAP | 34 |
| 3.4.4 Experiments Result and Analysis | 35 |
| 3.5 Summary | 36 |
| Chapter 4 Unsupervised Clustering and Speaker Recognition | 39 |
| 4.1 Introduction | 39 |
| 4.2 Review of Evaluation | 39 |
| 4.2.1 NIST2014 | 40 |
| 4.2.2 Whitening | 41 |
| 4.3 Experiments and Analysis of Unlabelled Data | 40 |
| 4.3.1 Database | 40 |
| 4.3.2 SVM Subsystem | 41 |
| 4.3.3 AHC Clustering | 45 |
| 4.3.4 AHC+PLDA Subsystem..... | 50 |
| 4.4 System Score Integration..... | 54 |
| 4.5 Summary | 56 |
| Chapter 5 Applications of Speaker Recognition | 59 |
| 5.1 Project Meaning | 59 |
| 5.2 Project Implementation | 59 |
| 5.3 Summary | 65 |
| Chapter 6 Conclusions and Future Works | 67 |
| References | 69 |
| Published Papers | 73 |
| Acknowledgement | 75 |

第一章 绪论

1.1 研究背景及意义

随着社会信息化的深入发展，人类生活的信息高度交互，特别是近期云计算和云存储的风靡，信息安全成为世界人民关注的重点问题。频发的互联网密码泄露，账号盗用事件已使人们对传统的身份安全认证方法失去了信任，社会迫切需要一种方便可靠的身份证鉴别技术来应对大众面临的恐慌与危机，于是生物特征识别技术走向了大众的视野，得到了大众的广泛关注。生物特征识别技术主要是指利用人类自身的生物特征信息进行身份认证的技术。能够进行生物识别的特征大致包括身体特征和行为特征两类^[1]，其中身体特征包括：指纹、静脉、掌形、脸形、虹膜、视网膜、DNA 信息等，行为特征包括：语音、签名、行走步态等^[2]。如今，人工智能和生物识别得到飞速发展，生物识别技术已经在很多领域有着实际应用，如指纹、人脸等识别技术在安防、考勤等领域广泛应用。

在众多的生物识别技术中，说话人识别^[3] ^[4] ^[5]具有特征采集设备成本低廉、可进行远程身份认证、隐蔽性好等优势，仅需要通过麦克风获取到语音信息，就能安全准确的确定本地或远程的身份信息。特别地，说话人识别技术已在不少司法社区矫正应用中得到采纳和实施。在云安全领域和智能终端设备上，个人隐私以及个人资料的安全问题亟待解决，说话人识别技术也被认为是一个非常好的应用解决方案。

说话人识别，在工程领域内又称声纹识别，它是指通过对说话人语音信号的分析处理，确认说话人身份信息的过程。说话人识别利用从语音波形中提取出反映说话人行为和生理特征的语音参数来完成自动识别说话人身份。它通过对说话人语音和数据库中已经登记的声纹作对比，对用户进行身份校验和鉴别。说话人识别与语音识别有着本质的区别，语音识别的目的在于确定说话人所讲述的内容，并不关心说话人身份的信息，而说话人识别的目的在于是确定说话人的身份信息，对说话人所讲述的具体内容不做深入研究。语音识别强调的是语义信息，对于不同说话人的差异应尽可能归一化；而说话人识别却恰恰相反，将语音中的语义信息尽可能平均化，突出包含在语音中的说话人的个性因素，把不同人之间的特征差异凸显出来。

说话人识别根据不同的任务类型可以分为两个范畴：说话人辨认(Speaker

Identification)和说话人确认(Speaker Verification)。在说话人辨认中，根据说话人训练目的地的不同，可以分为“闭集”和“开集”两种。闭集说话人辨认是基于测试语句的说话人一定是在训练集合内，而开集则认为待识别的说话人可能存在于训练的集合之内，也有可能存在于训练的集合之外。在具体应用中说话人辨认不要求说话人提供对应的说话人信息，系统会把待识别的说话人语音特征与预留的众多说话人特征模型相比较，从而确定说话人信息。因此，说话人辨认往往需要较长的时间开销，同时识别率也与集合内的模型数成反比。在说话人确认中，首先要求说话人提供对应的说话人信息，系统会验证待识别语音和言明身份者是否为同一人，系统把说话人语音特征与某个说话人的模型进行模式匹配得到匹配结果，因此该类识别性能会接近一个常数。此外，说话人确认和开集说话人辨认一样都通过一个阈值来判断测试说话人是否存在集合中。同时如果根据训练时的文本信息是否与测试时的文本信息一致，说话人识别又可以分为文本相关(Text-dependent)、文本无关(Text-independent)和文本提示(Text-prompted)三种^[6]。在日常信息处理中，特别是在安全监控和司法鉴定等领域，语音的文本信息通常是无法提前预知的，因此文本无关的说话人识别技术有更多的应用，也是我们研究的重点方向。

在说话人识别技术研究的历程中，说话人识别技术一直受困于语音处理领域内。机器学习的蓬勃发展给图像识别领域带来了技术的革新，使其性能获得了巨大提升，但语音由于其特征信息的特殊性，机器学习的理论成果不能得到充分应用，使得说话人识别技术在几十年间一直未搭上机器学习的“顺风车”，直到 i-vector 理论的提出，打破了两者之间的理论壁垒。因此本文研究的基于 i-vector 的说话人识别技术，具有很大的研究意义和实际应用价值。

下面介绍说话人识别技术的研究及国内外研究现状。

1.2 说话人识别技术研究及现状

利用说话人识别技术来分析识别说话人信息的工作可追溯到二战期间。1941 年，声谱仪在美国贝尔实验室的发明出现，开启了现代说话人识别技术的研究。二战结束后，贝尔实验室的物理学家 L·G·Kesta 受到美国司法局的委托，利用声谱仪通过观察来对语谱图匹配，首次提出来“声纹”的概念^[7]，由此开始了对声纹技术的科学的研究，他用对比语谱图的方法对文本固定的说话人语音进行辨认，为美国法院在取证方面做出了极大的贡献。

20世纪60年代，在数字信号处理技术的发展下，自动说话人识别技术开始萌芽。从60年代开始的基于模式匹配和统计方差分析方法，到后期的倒谱技术，从线性预测倒谱系数^[8]（LPCC）到Mel频率倒谱系数^{[9][10]}（MFCC）的概念，从识别参数提取方法的改进到模式匹配方法的革新，说话人识别技术不断更新发展。说话人识别技术逐步经历了从语音频谱和模板匹配^{[11][12]}，再到矢量量化^{[13][14][15]}（VQ）和动态时间规整技术^[16]（DTW）的发展，以及隐马尔可夫模型^{[17][18]}（HMM）的出现，在说话人识别领域中取得了出色的效果，成为说话人识别领域的核心技术，并且在不断改良的过程中，又提出了单个状态的HMM模型即高斯混合模型^{[19][20]}（GMM），GMM以其灵活性和鲁棒性，成为了说话人识别的主流技术。此外，人工神经网络也开始发展起来，它可以在一定程度上模仿人脑的功能，为说话人识别提供了一种新的方向。

1.2.1 说话人识别技术介绍

一个完整的说话人识别系统分为两个阶段：说话人训练和说话人识别。在说话人训练阶段，系统首先对提供的若干训练语音进行静音剔除和降噪处理，尽可能得到纯净有效的语音片段，然后再对这些语音提取对应的声学特征参数，根据系统的建模算法，得到说话人的特征模型。每个说话人对应的训练语音经过训练阶段后得到一个说话人模型。说话人识别系统的训练阶段如图1.1所示：

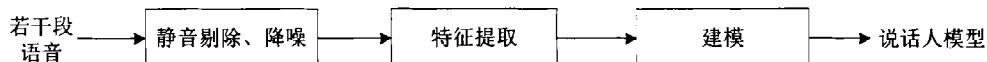


图 1.1 说话人识别系统训练阶段

在说话人识别系统的识别阶段，在这一阶段的主要任务是根据输入的测试说话人语音和其申明的说话人身份，对两者之间的模型进行模式匹配，根据匹配的结果判断测试语音是否和其申请的身份一致。在此过程中，对测试语音的处理与识别阶段对语音的处理过程是一致的，即需要对语音进行预处理后建模，得到测试语音对应的模型。说话人识别系统的识别阶段如图1.2所示：

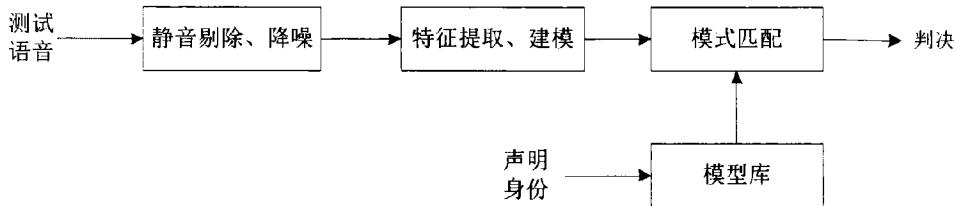


图 1.2 说话人识别系统识别阶段

在说话人识别系统中，说话人模型的训练过程是建模的过程，识别过程就是模式匹配的过程，由此不难看出，对说话人识别系统性能的提升，需要从建模和模式匹配这两个方向着手。

从说话人识别技术的发展来看，说话人识别模型经历了从矢量量化（VQ）、动态时间弯折（DTW）、高斯混合模型（GMM）、支持向量机^[21]（SVM）、到如今人工神经网络^[22]（ANN）的不断发展，说话人识别模型的发展史代表了说话人识别技术的进步史。

（1）矢量量化

矢量量化(VQ)来源于数字信号处理技术。应用在说话人识别的场景中，把待识别说话人的语音作为信号源，从该说话人的训练语音序列中提取的特征矢量组成一个用来表征该说话人的码本。在训练序列足够长的情况下，就可以认为说话人的个性特征已经包含在这个码本中。显而易见，对于 N 个说话人的识别系统，需要建立 N 个码本。要求这些码本在特征空间中相互之间不重叠。这便是基于 VQ 的说话人建模过程。在识别阶段，先从测试语音中提取出一组矢量 $V = \{v_1, v_2, \dots, v_t\}$ ；然后用系统中建立的 N 个码本分别对它们进行矢量量化，即识别这组矢量与特征空间中的哪一个码本的分布最吻合，吻合程度最高的码本就对应了系统识别结果。

（2）动态时间弯折

动态时间弯折(DTW) 作为一种非线性规整方法，同时结合了距离测度和时间规整计算，是说话人识别模式匹配中一种比较成熟的方法。在说话人训练过程中，特别是在固定语音文本的前提下，用这种文本固定的语音得到的特征序列作为对应说话人的特征。在识别阶段，要求说话人按照同样的文本内容进行发音，语音文件经过同样的特征提取后与模板进行匹配计算，最后根据得到的最小失真测度^[9]和阈值比较做出判决。

（3）高斯混合模型

GMM 模型^[23]不同于直接用语音的特征建立模型的方式，它是一种概率模型，建模

的依据来源于特征的概率分布情况。同时判决方式也发生了改变，它是根据似然得分来判决模型的相似性。每个人的语音声学特征空间可以用声学特征类表示，这些声学特征类代表着广义上的因素，GMM 之所以能够对说话人的特性进行有效的刻画，是因为它通过用多个高斯分布的组合来近似矢量的连续概率分布情况。在实际使用中，GMM-UBM^[23] 是一种最为常见和有效的模型，就是通过训练语音为每个人建立 GMM 模型，首先通过 EM 算法训练，估计一个高阶的 GMM 模型参数来刻画说话人的特征分布，训练语音未覆盖到的特征区域采用 UBM 的特征进行近似刻画，最后识别时计算所有测试语音特征与模型匹配的得分和，也就是计算最大似然率，作为识别结果。

（4）支持向量机

支持向量机(SVM)^[24]是一种目前比较流行的说话人建模技术，也是机器学习领域内重要的算法之一，具有较高的性能。和 GMM 模型不同，SVM 是一种基于区分性的模型，一种基于统计学习思想的小样本机器学习策略。作为一种分类器，SVM 可以对正负两类进行分类，凭借寻找最优的分类面来解决结构风险最小问题。SVM 具有较强的鲁棒性，与 GMM 相结合^{[25][26]}使用能够对系统识别性能做出较大提升。

（5）人工神经网络

人工神经网络（ANN）源于对人类大脑的仿生，模仿人脑的聚类和分类能力，是一种自适应系统。它是由大量的节点和节点之间的联接构成。不同的节点用来代表一种不同的输出函数，也称为激励函数^[27]。两个节点之间的联接代表加权值，也称为权重，以此来代表人工神经网络的记忆。通常下，一个神经网络是由多层的神经元组成的，上层的输出作为下层的输出^[28]，以此类推得到最终的输出结果。它被广泛应用于各种模式分类的应用中，通过自身的自适应和学习能力，对特征和模型不断优化更新，达到最优的输出结果。

在以上的几种说话人识别模型算法中，对语音的长度、文本及语音的信道等方面都具有一定的局限性，而在实际应用中短语音和跨信道问题的普遍存在，其中跨信道问题对说话人识别系统的性能带来的影响最大，这也是本文重点研究的问题。

1.2.2 说话人识别技术研究现状

在说话人识别的研究中，经典的 GMM-UBM（高斯混合模型-通用背景模型）系统虽然识别性能不错，但与实际应用所需求相比还具有一定差距，这是由于在跨信道问题、短语音问题、背景噪声问题、健康状况变化、说话人自身情感等方面对识别系统性

能有较为显著的影响。随着研究的深入，跨信道技术问题受到越来越多的科研工作者的重视，其中包括引入支持向量机(Super Vector Machine, SVM)这种区分性模型，挖掘语音数据中的未知高层特征参数^[29]，并实现各种得分规整技术如 C-norm^[30]、H-norm^[30]、Z-norm^[31]和 T-norm^[32]等。同时针对在说话人识别应用场景中的不同子系统，研究人员积极探索子系统之间得分的融合技术，比较常用的得分融合算法包括多子系统得分融合算法和分数非监督自适应算法等。实际应用中，由于电话系统语音采集设备和电话信道的多样性，不同信道之间的差异对电话语音系统识别的性能产生了很大的影响，因此如何实现复杂信道环境下说话人识别成为了研究热点。为了减小信道的差异对性能的干扰，很多学者提出了各种信道补偿算法，如说话人模型合成(Speaker Model Synthesis, SMS)^{[33] [34]}、特征映射(Feature Mapping, FM)^[35]、说话人识别联合因子分析(Joint Factor Analysis, JFA)^[36]以及扰动属性映射(Nuisance Attribute Projection, NAP)^{[37] [38]}。但是由于这些算法在运算中需要超高维矩阵的大量计算，在实际的应用场景中离不开高性能服务器的支撑，需要耗费巨大的计算资源。

在这样的背景下，基于 i-vector 的说话人识别技术应运而生。基于 i-vector 的识别技术不再严格区分说话人和信道的差异，将语音特征提取为一个低维的矢量矩阵，以此来表征说话人信息的差异性，在识别阶段只需计算矢量之间的余弦距离就可以作为相似性的评价标准，这就大大降低了计算复杂性和识别的时间成本。在 2014 年 NIST 评测中，i-vector 已经作为官方指定的说话人特征。在 i-vector 的基础上，Kenny 受到人脸识别中传统的线性鉴别分析^[39] (LDA) 的启发，提出了概率线性鉴别分析^{[40] [41]} (PLDA)，它是 LDA 的概率形式。PLDA 模型在说话人确认的过程中，在训练数据和测试数据信道不匹配的情况下，表现出了很强的鲁棒性。

随着多媒体时代的到来，说话人识别系统从成长逐渐走向成熟，从实验室走向实际应用。国内外声纹识别成绩显著，很多发达国家如美国、韩国、日本以及 ITT、Apple、Nuance、VoiceVault 等著名公司都为声纹识别技术的研发投入了巨资。其中 Nuance 公司提供的说话人验证解决方案使其成为全球说话人验证的领先者^[42]。随着客户流量从 PC 端向移动端的流动，移动互联网成为未来的发展趋势，因此，很多公司都纷纷加入到研究移动平台的安全防护上来，如 VoiceVault 开发的电话银行说话人识别系统以其优良的性能早已获得荷兰银行的青睐^[43]。另外借助于互联网开源工具^{[44][45]}，更多的研究人员投入到说话人识别的研究领域中来。国内相关企业的自主研发能力也一直紧跟着国际水平，北京得意音通与中国建设银行合作推出国内首个声纹识别的电话银行，厦门天聪公

司研发的文本相关的声纹识别技术已经应用于社区矫正对象的跟踪定位。

1.3 本文主要工作

本课题基于实际应用场景中的说话人识别系统，研究实现基于 i-vector 的算法和 PLDA 信道补偿算法，有效提高系统在实际应用中的整体性能和处理效率。同时针对现实应用中大量的无标注数据，本课题提供了一种实用高效的无监督聚类算法和流程。由于课题本身兼顾理论创新和实际应用价值，算法的时间复杂度也是本文关注的重点。

针对大量无标注数据，本文设计了基于 SVM 算法和基于 AHC 聚类的两套子系统，最后通过对聚类和基线两个子系统在得分域进行融合以达到最佳的性能。

1.4 本文结构安排

根据以上主要研究内容，本文的结构安排如下：

第一章 首先说明本文的研究背景及对现实应用的指导意义，其次介绍国内外说话人识别技术的研究历史和现状，最后给出本文研究的主要内容及结构安排。

第二章 介绍说话人识别技术算法的基本结构，并详细介绍了经典的 GMM-UBM 算法的具体方法和流程；而后在 GMM 的基础上引入 SVM 算法，介绍了基于 SVM 算法的说话人识别框架，讨论了原有算法存在的一些不足。

第三章 重点研究了基于 i-vector 的说话人识别系统。首先介绍了 i-vector 的基本思想和系统框架，给出了系统训练和验证的原理和流程，在理论上对系统的性能进行了分析评估；然后针对跨信道问题，着重研究了 PLDA 的信道补偿算法，对算法的原理和应用做出了详细阐述。

第四章 首先介绍了 NIST2014 评测情况，然后分析对 i-vector 进行白化处理对识别性能的影响；进而重点研究对无监督数据的分析和使用，在此基础上设计了基于 SVM 算法的子系统；同时引入无监督聚类，设计了基于 AHC 聚类和 PLDA 结合的子系统，最后在得分域进行系统融合，实现进一步提高系统识别性能的目标。

第五章 将本文研究的说话人识别系统应用到实际项目中，将实网数据引入到基于 i-vector 的系统中进行处理，验证课题研究的算法可行性、有效性和高效性，并分析实网数据中的语音对识别系统性能的影响。

第六章 对全文进行总结和工作展望，对实验和应用中遇到的问题进行了总结，并

指出目前存在的不足和有待进一步研究的工作。

第二章 说话人识别系统概述

2.1 引言

说话人识别系统的研究对象是说话人的语音，系统通过对语音文件进行静音剔除，降噪等处理后，提取蕴含在语音中的说话人个性特征来表征说话人的身份信息。在识别阶段，根据特征模型的模式匹配结果，系统做出说话人身份的判决。

在说话人识别系统中，语音的质量直接影响系统对说话人的特征提取，从而影响模型对说话人个性特征的刻画。因此语音的预处理和特征提取过程是整个识别系统工作的基础，也是保证系统识别性能的前提。

本章首先在说话人识别系统的基本结构上，介绍对语音文件进行预处理和特征提取的两个子模块；随后介绍经典的 GMM-UBM 算法的具体方法和流程；而后在 GMM 的基础上进而引出 SVM 算法，阐述基于 SVM 算法的说话人识别框架，最后分析算法存在的一些不足之处。

2.2 说话人识别系统的基本结构

以经典的基于 GMM-UBM 模型结构的说话人识别系统为例，详细介绍贯穿系统训练和识别过程中的有效语音检测和特征提取过程。一个完整的说话人识别系统的基本框架结构如图 2.1 所示：

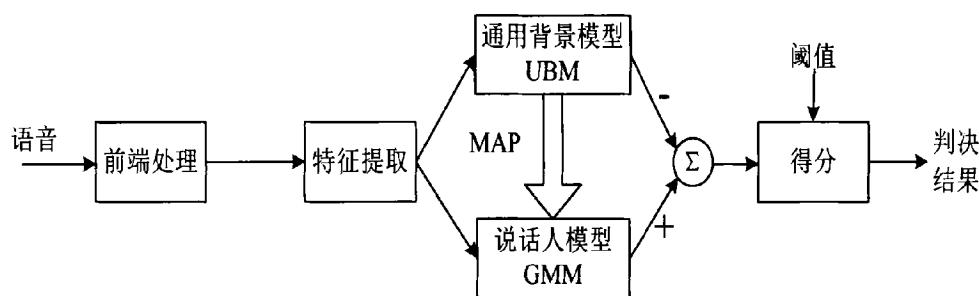


图 2.1 GMM-UBM 说话人识别系统基本结构图

2.2.1 有效语音检测

在说话人识别系统中，系统的研究对象是语音文件。但是在自然界中，一段语音文件中，除了有效的说话人语音信息外，还可能存在静音、彩铃、音乐及背景噪音等因素。静音中不包含任何说话人的特征信息，因此对说话人识别的效果是没有任何帮助的，但是语音文件中由于静音部分的存在，系统处理语音的时间成本相应增加。因此在语音信号处理的前端加入有效静音检测模块，剔除语音中的静音部分，对后续系统性能的提升有直接的意义。

有效语音检测^[46]（Voice Activity Detection, VAD）是指通过一定技术手段确定语音文件中有效的说话人语音部分。较为传统的有效语音检测方法是基于能量的语音检测，它是根据短时帧的能量大小来确定帧内语音段是否为有效语音片段。但是在对脉冲性噪声的处理上，这种处理方法会出现大量的误判，同时短时高能量噪声对说话人的识别性能会产生较大的影响，这在使用过程中具有很大的极限性。对于脉冲性噪声，基于滑动窗口的静音检测算法^[47]可以很好地解决这一问题。

基于滑动窗口的有效语音检测算法原理为：设计一个固定窗长的窗口，计算所有落在窗口内数据点的绝对值和，判断数据结果与预设阈值的大小，若低于预设阈值，则认为窗口内的语音段为静音，相反为有效语音。根据这样的原理，从语音数据的第一帧开始，窗口不断滑动，使语音中的每一帧都依次出现在窗口中，直到最后一帧处理结束。窗口从语音开头滑到末尾，就可获得整个语音的有效语音部分。同时，根据经验滑动窗口的窗长一般设为 200ms。根据上述原理介绍可知，在检测过程中，我们以一个窗长作为一个处理单位，如果窗长过长，就可能出现整个窗口判定为静音的情况，这会出现较多的误判，但是如果窗口过小，处理一段语音数据的时间成本就会增加，消耗更多的资源。因此选择经验值的窗口长度，在保证处理质量同时，兼顾了处理的效率。不过基于滑动窗口的有效语音检测也存在着不足之处，如果整体语音偏小，有效的语音片段被判定为静音的可能性相对较高。针对这样的问题，就需要引入分贝归一化等方法对语音的振幅进行整体的扩大。

经过有效语音检测处理后，语音文件内的所有语音片段均可以认为是有效的说话人语音信息，将剔除静音后的语音信息输入到系统特征提取模块，排除静音等因素对性能的干扰，确保语音特征对说话人身份信息的刻画。

2.2.2 特征提取

在模式识别领域内，待解决的问题可能是错综复杂的，也不可能把最原始的数据直接输入系统进行处理，我们需要从繁杂的原始信息中排除掉冗余信息，在减少后续数据量和计算量的同时，剩余的信息能够更好的表征问题的特性。从原始、具体的信息中获取抽象的、精简的信息，使得能够对原始信息进行最大程度的表征，这就是特征提取的目的。

语音特征参数的选择直接影响对说话人个性信息的表征，是影响识别性能关键因素。系统的识别率依赖于特征参数中具有区分性特征参数的提取。说话人识别其实就是对说话人的个性识别，每个人的发音器官构造都不同，发音习惯也是千差万别，对于人耳来说，某个人的声音越有个性，发音习惯越是特殊，就越容易把某个人的声音从众多声音当中区分开来，而计算机要像人一样区分出声音的个性特征那么首先就要选择好特征。

在理想情形下，声音特征应该具有以下特点^[19]：

- (1)能把说话人个体间差异充分体现出来，而对说话人本身的语音变化却保持稳定。
- (2)当语音信号受到信道噪音和外界噪音干扰时，其特征还能具有良好的鲁棒性。
- (3)不容易被冒充者模仿，特别是两个人的声音很相近时，更需要把个体间差异扩大化，把相似性缩小。
- (4)容易提取，方便计算，应该采用最经济最快速的提取方式来提取特征，不需要在特征提取环节占用太多的时间和内存消耗。

在说话人识别领域内较为常用的特征参数有线性预测倒谱系数(Linear Predictive Cepstrum Coefficient, LPCC)和 Mel 频率倒谱系数(Mel-Frequency Cepstrum Coefficient, MFCC)^{[9] [10]}。语音信号的相邻两帧之间存在着较大的相关性，往往需要在静态的倒谱中加入动态信息来强化特征表示，如添加倒谱的一阶差分或二阶差分^[50]，然后再对特征进行倒谱均值减和倒谱方差归一化的处理，以消除传输信道的影响和信道带来的偏移误差。

LPC 是语音处理中一种重要的技术。它可以用很少的参数正确而又有效的表现语音波形及其频谱的性质，并且计算效率高，便于应用。利用 LPC 系数推导出倒谱系数，即 LPCC，主要反映了声道的响应，而且只需要较少量的倒谱系数就能较好的刻画语音的共振峰特性。但是 LPCC 也存在缺陷，人耳的实际听觉频率特性是非线性的，而 LPCC 是一种时间序列的线性预测模型，这与语音实际不符。

人耳之所以在嘈杂的环境中以及恶劣的情况下仍能够正常地区分出各种语音，这与耳蜗的作用是分不开的。耳蜗本质上就相当于是一个滤波器组，而耳蜗的滤波有一个特点，就是对 1000Hz 以下的语音信号，它呈现出线性尺度，而对于 1000Hz 以上的语音信号，它呈现出对数尺度，即人耳对低频信号敏感，对高频信号不敏感。于是，人们又找到了类似于人耳听觉特性的非线性参数，即 Mel 倒谱系数。图 2.2 就是 Mel 频率与声音物理频率之间的关系：

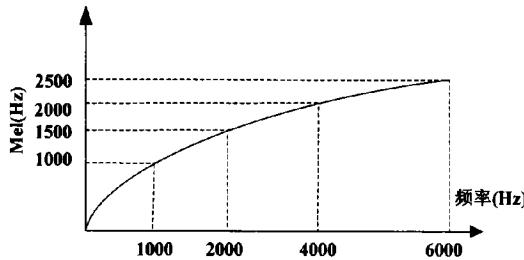


图 2.2 Mel 频率与物理频率的对应关系

我们可以用一个近似公式来表示：

$$F_{Mel} = 2595 \lg\left(1 + \frac{f_{Hz}}{700}\right) \quad (2-1)$$

MFCC 参数是在没有任何前提假设的基础上充分模仿了人耳的听觉特性，而且具有良好的抗噪能力和识别性能。因此，它在说话人识别和信号处理领域广为应用。

求解 MFCC 参数的过程可以分为以下几个步骤：

- (1) 将语音信号模数转化，预加重，分帧和加窗处理，再经快速傅立叶变换(FFT)得到其频域信号。
- (2) 把频域信号通过 D 个 Mel 滤波器组，得到其 D 维的 Mel 频谱，再求频谱平方计算能量谱。
- (3) 对能量谱取对数。
- (4) 再把求得的对数能量谱经离散余弦(DCT)变换得到 MFCC 参数。
- (5) 而经过 DCT 变换求得的 MFCC 是静态特征，一般我们还要加入静态特征的一阶或二阶差分。

我们可将上述的过程通过流程图 2.3 描述：

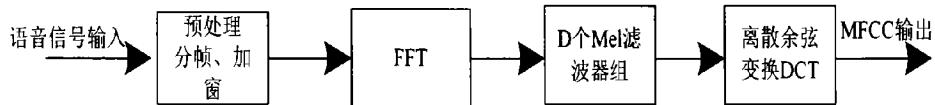


图 2.3 Mel 参数提取流程示意图

说话人识别系统中通常使用 MFCC 参数，本文 D 取 24，特征参数的阶数设为 16，并取其一阶动态特征，总共 32 维。

信道产生的平稳卷积噪音会对语音部分产生干扰，通常在 MFCC 提取完之后加入倒谱均值减（CMS）的处理，CMS 利用噪声频谱相对独立，且满足一定的分布规律的特点，利用噪声特征的均值对带噪语音进行特征补偿，具体计算过程如下：

$$C_d'(t) = C_d(t) - \frac{1}{N} \sum_{i=1}^N C_d(i), \quad d = 1, 2, \dots, D \quad (2-2)$$

式中， $C_d(t)$ 表示第 t 帧的第 d 维分量， D 是特征的维度， N 是总长度。

为了消除信道带来的偏移误差，可以在系统中加入倒谱方差归一化（CVN）具体公式如下：

$$C_d'(t) = \frac{C_d(t)}{\sigma_d}, \quad d = 1, 2, \dots, D \quad (2-3)$$

式中， σ_d 是倒谱特征计算中的方差的第 d 维系数。

我们把 CMS 和 CVN 合并在一起，就可以得到：

$$C_d'(t) = \frac{C_d(t) - \frac{1}{N} \sum_{i=1}^N C_d(i)}{\sigma_d}, \quad d = 1, 2, \dots, D \quad (2-4)$$

对 MFCC 求取一阶差分特征，得到以下公式：

$$\Delta C_d'(t) = \frac{\sum_{n=1}^N n^2 \frac{(C_d'(t+n) - C_d'(t-n))}{2n}}{\sum_{n=1}^N n^2} \quad (2-5)$$

式中， $\Delta C_d'(t)$ 表示第 t 帧的 MFCC 动态差分， n 表示差分窗口长度。经过倒谱均值减以及倒谱方差归一化处理后，特征参数的信道鲁棒性得到提升。

2.3 GMM-UBM 说话人识别系统

上一节主要介绍了说话人识别系统的基本结构，并讨论了在系统前端对输入语音信号进行处理的有效语音检测模块和特征提取过程，本节将在此基础上详细分析在说话人识别领域经典的 GMM-UBM 模型。

2.3.1 GMM 简介

高斯混合模型(GMM)来源于隐马尔科夫模型(HMM)，它是只有一个状态的 HMM，不同的是，不像 HMM 需要各态历经，不再考虑 HMM 中声学特征的时序性，这也大大减小了建模的计算量。GMM 因其简单灵活而高效的特点，在说话人识别应用中一直被人们沿用至今。首先它是一种概率统计的模型，是由若干个高斯概率密度函数的线性组合而成，GMM 一般被用于文本无关的说话人识别系统，但在实验中也发现 GMM 也具有文本相关的效果。它有两大优点：第一，说话人的声学特征参数可以认为是发不同音时的声学特征在特征空间的集合而成。第二，多个高斯概率函数线性组合可以拟合任意分布，因此它可以拟合任意形式语音特征参数的概率分布^[20]。基于这两方面的优点，我们就可以从不同的语音参数特征分布中找出不同的说话人。图 2.4 就是 GMM 模型构成的例子，该 GMM 模型是由 4 个高斯概率密度函数的线性组合构成。

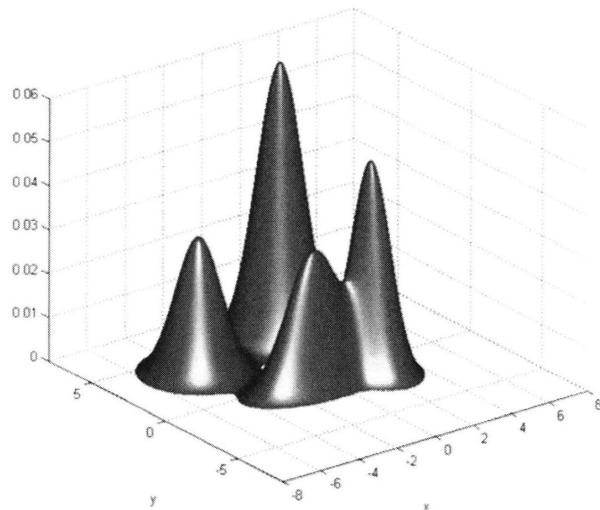


图 2.4 GMM 模型构成

对于 M 阶 GMM 的高斯分量可以如下描述：

$$P(x | \lambda) = \sum_{i=1}^M P(x, i | \lambda) = \sum_{i=1}^M c_i P(x | i, \lambda) \quad (2-6)$$

式中， λ 为 GMM 模型； x 为声学特征，维度为 K ； i 为高斯分量的序列， M 阶 GMM 就是有 M 个高斯密度函数； c_i 为第 i 个分量的高斯函数权重，其值所对应的为密度函数 i 的先验概率，且有：

$$\sum_{i=1}^M c_i = 1 \quad (2-7)$$

$P(x | i, \lambda)$ 为高斯函数的混合分量，对应为密度函数 i 的观察概率密度函数，通常用 K 维的单高斯分布函数，用下式表示：

$$P(x | i, \lambda) = \frac{1}{(2\pi)^{K/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)}{2}\right\} \quad (2-8)$$

这里， Σ_i 和 μ_i 分别为高斯函数的协方差矩阵和均值参数； $i = 1, 2, \dots, M$ 。

M 阶 GMM 是用 M 个高斯函数分布的线性组合来描述，也就是 GMM 参数集 λ 由均值参数、协方差矩阵以及高斯函数分量的权重组。可表示成如下三元组的形式：

$$\lambda = \{c_i, \mu_i, \Sigma_i; (i = 1, \dots, M)\} \quad (2-9)$$

Σ_i 可以是普通矩阵，也可以是对角矩阵。一般都取使系统算法简单、性能也好的对角矩阵。即

$$\Sigma_i = \text{diag} \{ \sigma_{i0}^2, \sigma_{i1}^2, \dots, \sigma_{iK-1}^2 \} \quad (2-10)$$

$\sigma_{ik}^2 (k = 0, 1, \dots, K-1)$ 为 GMM 第 i 个分量所对应特征矢量的第 k 维分量的方差。

将两式合并可以得到：

$$P(x | i, \lambda) = \prod_{k=0}^{K-1} \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left\{-\frac{(x_k - \mu_{ik})^2}{2\sigma_{ik}^2}\right\} \quad (2-11)$$

x_k 和 μ_{ik} 分别是矢量 X 和矢量 μ_i 的第 k 个分量。

均值决定了高斯的位置，方差决定高斯的分布，权重决定了高斯分布的幅度。因此对于任意一个分布来讲，只要找到合适的均值、方差和权重就能够进行拟合。寻找这样的模型参数 λ 可以采用 EM 算法^[51]进行迭代。对于 EM 算法，拟合精度与训练数据为正

比关系，但是在应用中通常不具备大批量训练数据的条件，因此引入了通用背景模型（UBM）的概念，认为经过数据充分训练的 UBM 足以表征所有说话人的特征信息。

2.3.2 GMM-UBM 的训练

为了解决训练不充分，且信道，环境等不匹配的问题，D.A.Reynolds 提出来 UBM 的概念^[23]。UBM 就是用大量的说话人语音训练出的一个高阶高斯模型，它包含了大量说话人的特征信息。在应用中把 GMM 和 UBM 结合起来，用 GMM 来表征说话人个性特征，UBM 来表征说话人的共性特征。对于一个具体的说话人 GMM 模型，它是在 UBM 上由说话人的语音通过最大后验概率（MAP）^[12]自适应得到。训练过程等价于对 UBM 中的高斯进行修正的过程，同时由于训练语音每一帧都会在修正的过程中体现出来，因此自适应得到的 GMM 模型能够充分表征说话人的个性特征。图 2.5 为在 UBM 上训练语音自适应过程的示意图：

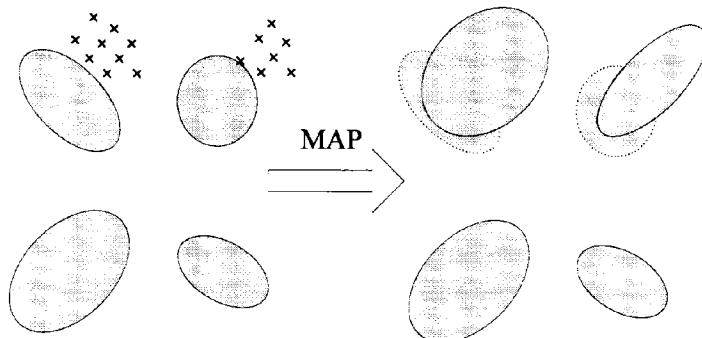


图 2.5 UBM 的 GMM 自适应过程示意图

在 GMM-UBM 的模型下，通过 MAP 自适应只是修改了 UBM 和目标说话人特征中相近部分的高斯分布的描述，突出目标说话人的个性特征所在；然而那些与目标说话人特征分布较远的部分不作变动，表明目标说话人和冒充者的共同特性。实验表明，训练数据的时间越长，目标说话人的特征分布也就越可靠，通过 MAP 修正后模型越接近目标说话人的分布而远离 UBM。实验也证实，只修正均值时系统效果能够达到最好，即目标说话人模型的权重和方差与 UBM 相同，均值不同。运用 MAP 算法，在 GMM 与 UBM 之间的高斯概率密度函数建立了相互对应的关系，这种关系有效的抵消了声音因素的影响，凸显目标说话人的个性特点，因此 GMM-UBM 系统性能要优于 GMM。

从理论上讲，训练 UBM 的说话人数据越多，包含信道情况越全面，涵盖说话人越

广泛，系统的性能会越好。

2.3.3 GMM-UBM 似然比得分

说话人确认系统，就是先让待测说话人言明身份，再作说话人识别。我们事先假设有一个待确认的说话人 S ，对于一个测试的语音 X ，定义：

H_0 ：测试语音 X 来自目标说话人 S ；

H_1 ：测试语音 X 不是来自目标说话人 S ；

我们通过给定一个阈值 θ ，和一个对数似然比来评判属于上述中的哪一种情况。对数似然比定义为^{[45][46]}：

$$S(X) = \log\{P(X|H_0)/P(X|H_1)\} \quad (2-12)$$

当 $S(X)$ 大于等于 θ 时，表示测试语音是属于说话人 S ；

当 $S(X)$ 小于 θ 时，表示测试语音不是说话人 S 。

因此，在确认系统中，设定恰当的阈值是一个关键因素，而关于该阈值至今也没有特定的计算方法，通常靠经验和具体测试来调整。

在上述公式中，我们可以把计算两个概率 $P(X|H_0)$ 和 $P(X|H_1)$ 认为是计算待测语音分别与目标说话人模型匹配的概率以及与冒充者模型匹配的概率。

因此公式 $S(X)$ 可以写成：

$$S(X) = \log\{P(X|H_0)/P(X|H_1)\} = L(X|\lambda_s) - L(X|\lambda_{ubm}) \quad (2-13)$$

在应用中， λ_{ubm} 这个模型 UBM 通常具有很大的高斯混合密度数量，需要很多不同说话人的数据来训练得到。

由公式 2-13 可知，在相减的过程中，模型间相同的部分相互消去，凸显出了说话人个性部分，同时信道噪音和背景的等干扰因素被抑制，较 GMM 模型，提升了系统的鲁棒性。

2.4 GMM-SVM 说话人识别系统

上一节介绍的 GMM-UBM 模型在说话人特征空间的刻画和鲁棒性方面虽然有不错的性能，但是 GMM-UBM 在训练时只关注自身语音特征分布的情况，对于相似的说话人和说话人之间的特征差异未做考虑。在训练阶段，如果增大目标模板的似然概率，与说话人相似的竞争模型的似然概率随之增大，虽然提高了自身模型识别成功率，但是同

时对不同说话人的区分性下降，增大了冒充成功的风险，导致系统整体性能下降。因此，研究者将机器学习领域内经典的 SVM 方法引进说话人识别领域，构建了基于 GMM-SVM 的说话人识别系统。

2.4.1 SVM 简介

SVM (Support Vector Machine)^[24]是一种基于统计学习理论的有监督的二元分类器。可以有效地从一组有监督的样本中找到最佳的分界面，把样本中的正样本点和负样本点线性区分。说话人识别系统核心问题就是分类，因此作为一个分类器，SVM 可以有效的处理说话人识别的问题，

将 SVM 运用到说话人识别中，需要较多的正样本点和大量负样本点，这就必然导致有大量的目标语音和冒充语音需要处理。而在说话人识别应用中，往往冒充者的语音要远远多于目标语音，容易造成正负样本之间的数量失衡。其次，语音特征的主流提取方法，如使用 MFCC 算法提取出的说话人特征参数中，说话人的特征信息和文本信息，语义信息等夹杂在一起，说话人的个性特征区分性不强，单纯使用 SVM 算法很难准确的区分。

为解决上述问题，研究人员进一步提出将 GMM-UBM 模型和 SVM 这种区分性模型结合起来，构建基于 GMM-SVM 的说话人识别系统。GMM-SVM^[21]系统的前端部分采用 GMM-UBM 模型的前端，将语音的频谱特征作为输入量，特征改为用高斯超向量（Gaussian Supper Vector, GSV）^[23]表征，然后再使用 SVM 基于特征分类。

2.4.2 GMM-SVM 说话人识别系统框架

在 GMM-SVM 说话人识别系统中，高斯超矢量 GSV 是由 GMM-UBM 中得到的说话人模型的各个混合分量的均值向量拼接而成，它是一个高维的超级向量。假设 GMM 系统中 GMM 的高斯分量为 M，输入为 D 维特征向量，则 GSV 向量的维度为 M*D。GSV 的获取过程如图 2.6 所示：

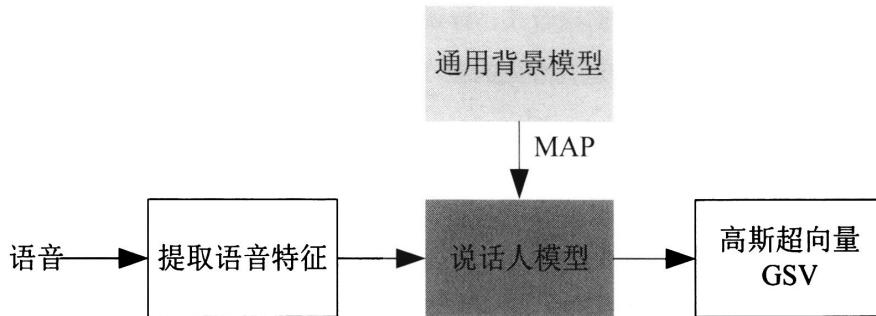


图 2.6 GSV 超向量的获取过程

GSV 这个超级向量作为话人的特征向量输入到 SVM 中，再由 SVM 算法进行特征的分类判别。

GMM-SVM 说话人识别系统的系统流程框图如图 2.7 所示：

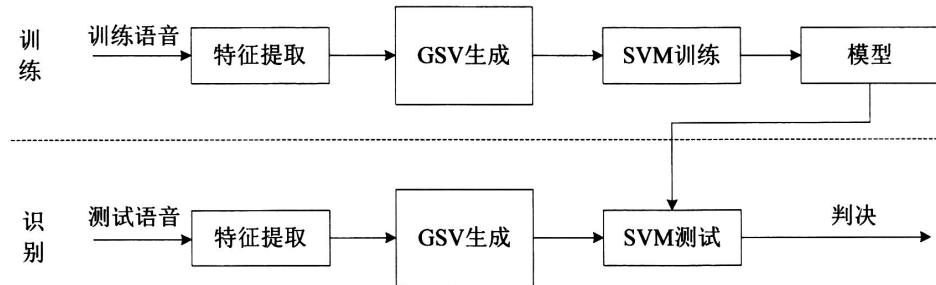


图 2.7 GMM-SVM 说话人识别系统框图

基于高斯超向量 GSV 构造的 GMM-SVM 系统，SVM 在训练阶段需要说话人的正例样本和负例样本。负例样本来源于大量的冒充者的语音在 UBM 上通过 MAP 自适应算法获得的高斯超向量；在训练说话人模型时，若有多段训练语音，则对应会有多个正例高斯超向量，将正例样本与负例样本一起作为 SVM（使用线性核或 KL 散度核）的输入，训练一个超平面以表征目标话者的模型；在识别时，将测试语音从 UBM 中通过 MAP 自适应算法得到 GSV，以此作为 SVM 输入再与说话人模型对比识别（数学上就是计算 $w^T x_{test} + b$ 的过程），最后再与阈值比较做出判断。

阈值的确定来源于目标说话人和冒充者在目标说话人的 SVM 模型上的得分分布情况。在 SVM 分类中，由于训练语音往往较少，因此正例样本点也较少，因此分类面不能仅仅由正例决定，对应的阈值也不能仅有目标说话人的语音确定。想要获得通用的阈

值，就需要考察大量非目标说话人的语音在目标说话人的模型上的得分分布以及在各自模型上的得分分布情况。只有综合这几方面的分布情况，才能确定一个合适通用的阈值门限。

2.4.3 SVM 系统在说话人识别应用中面临的问题

在 GMM-SVM 说话人识别系统中，融合了 GMM 中对说话人语音特征的刻画能力强和 SVM 区分性强的优势，将说话人的特征从低维空间扩展到高维空间，从而更易于区分，也成为了说话人识别领域内一个重要的处理方法。

虽然 GMM-SVM 算法相对于 GMM-UBM 算法在识别性能上明显提升，在说话人识别领域也占有一席之地，但是依然存在着诸多问题：

(1) 对短语音的识别性能急剧恶化。对于说话人特征的提取一般采用的 MFCC 算法，GMM-SVM 说话人识别系统也不例外，但是 MFCC 提取的特征参数中既包含说话人的特征信息，又包含有语义信息。说话人的个性特征是依赖于说话人语音对 UBM 的自适应得到，当语音较短时，自适应得到的参数相对于 UBM 的参数偏移不明显，无法保证对说话人个性的准确刻画。这样有可能会使拼接出的 GSV 在某些维度上对说话人的表征是不准确的，如果训练语音和测试语音的语义不同，会造成两个超向量中对应维度对说话人的刻画不匹配，进而不能准确反映正确的匹配情况。

(2) 跨信道问题。实用环境的复杂性让说话人识别的跨信道问题成为一个重要的研究方向，信道的差异经过特征提取后和说话人的个人特征混杂在一起，系统无法区分不匹配的原因来自于信道差异还是个人特征的不同，造成系统误判率上升。在应用场景下，跨信道问题是无法避免的情况，虽然 GMM-SVM 系统在原有基础上，加入有害因子投影 (NAP) 算法，在一定程度上提升系统的信道鲁棒性，但是以 NAP 算法的计算量添加到计算复杂度本来就很高的 GMM-SVM 系统上，无疑于雪上加霜。

(3) 计算复杂度高，在处理大规模说话人识别应用中效率下降。在实际应用中，语音数据的数据量动辄十万，百万级，对于 GMM-SVM 系统，无论是 GMM 还是 SVM，其计算复杂度，合并与分类的数量均与模型个数为正比例关系。在应用工程中，时间成本的开销也是重要的考虑因素，随着数量增加识别速度和效率急剧下降的 GMM-SVM 已不能满足项目对时间的要求。

由此可知，说话人识别系统应用的重点在于提高系统鲁棒性和高效性，克服应用中的短语音问题，在算法层面减小环境和信道对性能的影响，其次在保证系统性能的情况下

下，提升系统处理的效率。

2.5 本章小结

本章首先介绍了说话人识别系统的整体架构，给出了说话人识别系统的框架图，然后详细介绍了对输入语音信号进行有效语音检测和特征提取过程，在此基础上阐述了经典的 GMM-UBM 算法中训练和识别得分的理论依据，而后在 GMM 的基础上引入 SVM 算法，分析介绍了 SVM 的工作原理，通过对基于 GMM-SVM 算法说话人识别框架的分析，讲述了系统的工作原理和流程，最后讨论了算法中存在的一些不足之处。

此页不缺內容

第三章 基于 i-vector 的说话人识别系统

3.1 引言

上一章介绍了说话人识别领域内经典的 GMM-UBM 模型和 GMM-SVM 模型。两种算法在说话人识别中都取得了不错的性能，但是在面对跨信道问题和应用中大规模数据时，算分的性能和效率还达不到应用的要求。为了提高系统在信道失配环境下的鲁棒性，Kenny 提出了联合因子分析（JFA）^[52]技术，JFA 在表现出了优秀的跨信道能力，但同时增加了大量的计算，在实际应用中难以推广。在 JFA 的基础上，Dehak 提出了鉴别性向量 i-vector 模型^[53]。i-vector 模型的提出，打破了语音特征和机器学习之间的技术壁垒，迅速成为说话人识别领域的主流技术。由于 i-vector 的极佳性能，成为美国国家标准与技术协会（National Institute of Standards and Technology, NIST）2014 评测官方指定的说话人特征。

基于 i-vector 的说话人识别系统，相对于 JFA 计算量大大降低，这就让说话人识别应对大规模数据成为可能。在对信道鲁棒性的提升方面，i-vector 本身具有不错的跨信道能力，同时 PLDA^[41]信道补偿算法的引入，使得系统鲁棒性得到进一步提高。

本章内容安排如下：首先介绍 i-vector 模型的基本思想和框架；然后阐述基于 i-vector 模型说话人识别系统的原理和流程；接着分析 i-vector 在识别性能和效率上的性能；最后引入基于 i-vector 的 PLDA 信道补偿算法，并介绍 PLDA 模型在训练阶段和打分阶段的理论依据和应用。

3.2 基本思想

在 JFA 算法中，将说话人的超矢量空间划分为：本征信道(EigenChannel)空间、本征音(EigenVoice)空间和残差空间（残差空间一般都非常小，可忽略）。通过对本征信道空间说话人因子的处理，达到提升系统跨信道的能力。在 i-vector 模型中，使用单一的空间来代替 JFA 中本征信道空间和本征音空间，我们称之为全局差异空间(Total Variability Space)^[53]。说话人间的差异和信道的差异都包含在这一个空间中。在 i-vector 中，说话人信息和信道信息同时对 GMM 超空间矢量产生影响，两者不做严格区分。之所以把两

者的影响不作区分，采用新的空间模型，是因为在实验中发现JFA中的本征信道空间虽然是不同的信道建模而来，但其实已经夹杂了说话人的特征信息^[54]。

i-vector的基本思想表示如下：

$$M = m + Tw \quad (3-1)$$

其中， m 是说话者人无关且信道无关的均值超向量，通常由UBM的均值向量拼接而成； T 是一个低秩的矩阵，表示全局差异空间；而 w 则是服从标准正态分布的随机向量，我们称之为identity-vector，简称i-vector。

在这个过程中，将包含说话人和信道信息GMM均值超向量向低秩的全局差异空间上进行投影，可以得到一个低维矢量 w ，维度一般为400-600，它仅含有说话人的信息，表征了说话人的个性特征。

3.3 基于 i-vector 模型的说话人识别系统

根据对i-vector模型原理的介绍，一个基于i-vector模型的说话人识别系统有两个关键方面，即全局差异空间的估计和i-vector的估计。一个刻画准确的全局差异空间可以保证GMM的均值超向量在投影后对说话人的信息和信道信息区分性更加明显，能够得到一个相对纯净的表征说话人特征的i-vector。每个说话人的均值超向量经过投影后得到对应的i-vector，最后对这些i-vector根据余弦距离进行打分，最终得到判定结果。

基于i-vector的说话人识别系统结构如图3.1所示：

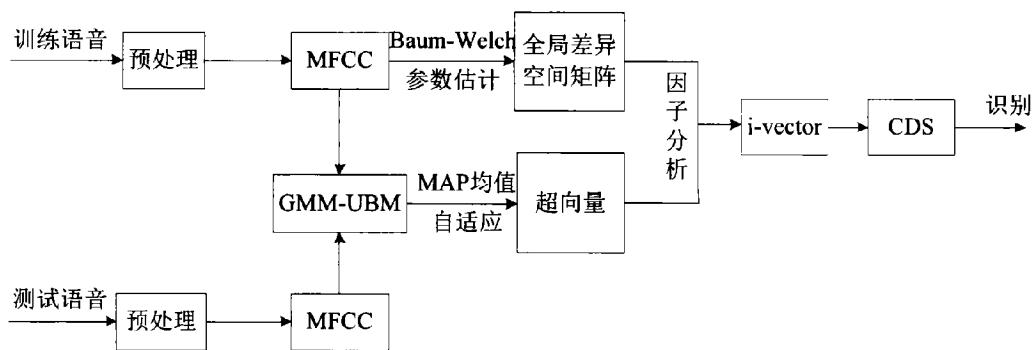


图 3.1 基于i-vector的说话人识别系统框图

3.3.1 全局差异空间的估计

训练全局差异空间需要大量的语音文件，而且在对语音的使用上有所不同。全局差异空间认为所有给定的语音数据都是来自于不同的说话人，即便是一个说话人的多段语音也同样认为是来自于不同人。

对训练全局差异空间 T 矩阵的估计，首先要计算每个说话人对应的 Baum-Welch^[55] 统计量。说话人 s 的语音段 h 表示为 $x_t = \{x_1, x_2, \dots, x_L\}$ ，用 λ 表示 M 阶的 UBM 模型。

$$N_{j,h}(s) = \sum P(j | x_t, \lambda) \quad (3-2)$$

$$F_{j,h}(s) = \sum P(j | x_t, \lambda) x_t \quad (3-3)$$

其中，变量 j 是 UBM 中高斯个数变量， $j=1, 2, \dots, M$ ， $P(j | x_t, \lambda)$ 为语音中第 x_t 帧特征对于 UBM 中第 j 个高斯分量的后验概率。

在后续计算中会需要 Baum-Welch 的一阶中心统计量 $\tilde{F}_{j,h}(s)$ 。 m_j 表示 UBM 模型中第 j 个高斯分量的均值。

$$\tilde{F}_{j,h}(s) = \sum_{t=1}^L P(j | x_t, \lambda) (x_t - m_j) \quad (3-4)$$

即：

$$\tilde{F}_{j,h}(s) = F_{j,h}(s) - N_{j,h}(s) m_j \quad (3-5)$$

为了计算方便，可以将统计量扩展为矩阵形式：

$$NN(s) = \begin{pmatrix} N_1(s) & & 0 \\ & \dots & \\ 0 & & N_M(s) \end{pmatrix} \quad (3-6)$$

$$FF(s) = \begin{pmatrix} \tilde{F}_1(s) \\ \dots \\ \tilde{F}_M(s) \end{pmatrix} \quad (3-7)$$

Baum-Welch 的统计量计算完成后，就可以进行 T 矩阵的训练，采用的是 EM 迭代算法。

E 步骤：在训练之前，对 T 矩阵进行随机初始化。然后计算说话人因子的方差和均值：

$$l_T(s) = I + T^T \Sigma^{-1} NN(s)T \quad (3-8)$$

$$\bar{y}(s) = l_T^{-1}(s)T^T \Sigma^{-1} FF(s) \quad (3-9)$$

对于全局差异空间的总体因子其后验概率分布满足均值为 $\bar{y}(s)$ ，协方差矩阵为 $l_T(s)$ 的高斯分布。

M 步骤：进行最大似然值重估。

通过训练累计所有语音统计量：

$$N_c = \sum_s N_c(s) \quad (3-10)$$

$$A_c = \sum_s N_c(s) l_T^{-1}(s) \quad (3-11)$$

$$C = \sum_s FF(s) \bar{y}(s)^T \quad (3-12)$$

对于所有语音，统计上述统计量后，需要对全局差异矩阵进行更新，更新公式为：

$$V = \begin{pmatrix} V_1 \\ \vdots \\ V_c \end{pmatrix} = \begin{pmatrix} A_1^{-1} C_1^T \\ \vdots \\ A_c^{-1} C_c^T \end{pmatrix} \quad (3-13)$$

此时判断 EM 算法迭代次数是否已达到初始设定迭代次数，若没有，则重新返回到 E 步骤对说话人因子的均值和方差进行统计计算，M 步骤再次进行最大似然估计，对 T 矩阵继续更新，直至到达设定迭代次数。根据经验值，迭代次数一般设定在 10 次左右，因为每次迭代过程需要大量的运算，耗时较长，经过 10 次迭代后的 T 矩阵已较为收敛。

至此，全局差异空间矩阵 T 训练完毕。

3.3.2 i-vector 的估计

训练完全局差异空间矩阵 T 后，就可以进行 i-vector 的提取。根据 Baum-Welch 统计量得到其后验概率分布情况，其后验概率分布满足高斯分布。 w 是服从标准正态分布的随机向量，因此对用 i-vector 的估计值 w 就是高斯分布的均值矢量。

对 i-vector 的估计过程为：

首先，根据目标说话人语音，计算说话人对应的 Baum-Welch 统计量；

其次，读取全局差异空间矩阵 T ；

最后，根据公式 3.14 计算说话人对应的 i-vector 的估计值 w 。

$$E[w_{s,h}] = (I + T^T \sum^{-1} N_h(s)T)^{-1} T^T \sum^{-1} \tilde{F}_h(s) \quad (3-14)$$

其中， Σ 是 UBM 的协方差矩阵。如果目标说话人的语音有 h 条，则通过上述公式计算可以得到 h 条 i-vector 矢量。一般情况下 i-vector 的维度在 400-600 之间。

至此，i-vector 的提取工作已经完成。

3.3.3 系统测试打分

系统得到两个不同说话人的 i-vector 矢量后，对于两个矢量之间的相似程度，经常采用两个矢量之间的余弦距离^[56]来衡量。例如，特征矢量 w_1 和 w_2 之间的相似程度可以表示为：

$$k(w_1, w_2) = \frac{\langle w_1, w_2 \rangle}{\|w_1\| \|w_2\|} \quad (3-15)$$

需要注意的是，只有在同一个全局差异空间矩阵 T 上投影得到 i-vector 矢量，考察相互之间的相似度才是有意义的。基于不同的 T 矩阵提取的矢量特征相互之间不具有可比性。由公式 3.15 可知，两个矢量之间只考虑其夹角，这样不仅能够有益于设定统一的阈值，同时对系统的鲁棒性也有一定提升。

说话人的测试过程就是考察目标说话人和测试语音之间矢量特征相似程度的过程。在 i-vector 模型中，i-vector 矢量就代表了说话人的特征，因此只需计算目标说话人矢量 w_{target} 与测试语音矢量 w_{test} 之间的余弦距离。

$$score(w_{target}, w_{test}) = \frac{\langle w_{target}, w_{test} \rangle}{\|w_{target}\| \|w_{test}\|} > \theta \quad (3-16)$$

矢量之间的余弦距离与阈值 θ 进行比较，大于阈值，则认为测试语音属于目标说话人，否则，认为不是同一个人。

相对于 SVM 和 JFA 等模型在测试阶段考察相似性的模式匹配算法，余弦距离算法使打分过程更加快速和简化。

3.3.4 系统性能分析评估

联合因子分析（Joint Factor Analysis, JFA）算法建立在 GMM 和 SVM 两种模型的理论基础之上，同时也充分利用了生成性模型和区分性模型的优点。它克服了 GMM 模型区分性不足的问题，同时也解决了 SVM 面对大规模背景训练数据处理困难的问题。但是 JFA 对训练数据要求非常苛刻，同时需要估计大量参数，计算量非常大，尽管从理论上效果十分诱人，但是在应用中实现难度较大。因此实验采用基于本征信道分析（Eigen Channel），又称隐藏因子分析（Latent Factor Analysis, LFA）算法。LFA 是一种简化因子分析算法，它将超向量在经过简化后进行分解，转换为说话人和信道信息两个超向量之和，这就对计算难度大大简化。

i-vector 模型作为 JFA 模型的优化形式，将原有的 GSV 通过在全局差异空间中的投影过程，将说话人信息和信道信息有效的区分，同时低维度的全局差异矩阵和 i-vector，极大程度上的降低了识别阶段的计算量，提升了系统的性能和效率。

为了对系统性能有更直观的认识，考察基于 i-vector 的说话人识别系统相对于经典说话人识别系统的性能表现，设计实验对 4 种说话人识别算法进行性能比较，实验数据来源于电话实网数据，语音信道不明，其中确认测试：1166 次；冒充测试 31569 次。实验结果如图 3.2 所示：

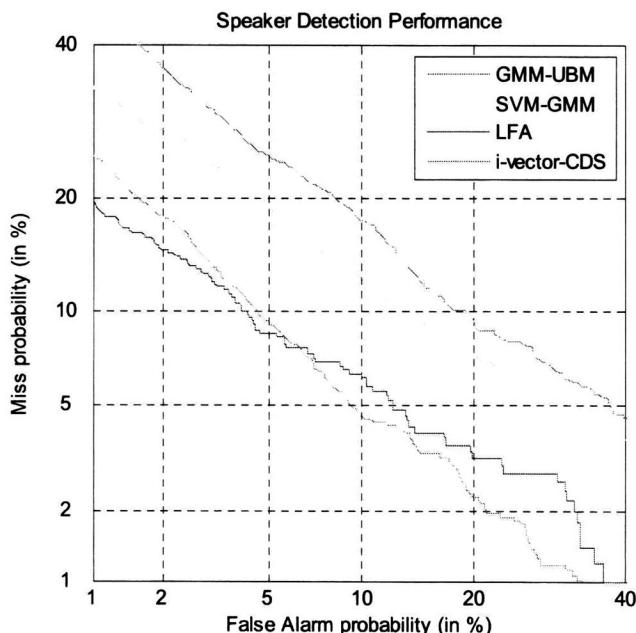


图 3.2 不同说话人识别算法性能对比曲线

在实际应用中，在考虑算法的性能的同时，算法的效率同样也至关重要，为了对算法的效率有直观的了解，设计了一组实验考察不同算法在同样的硬件条件下，处理相同数据的耗时情况。因为在实际应用场景中，主要是对说话人识别系统的识别阶段的应用，因此实验只是考察算法在识别过程中的耗时情况，训练阶段在应用中基本都是离线处理，因此训练阶段的耗时不在本次实验的考察范围内。实验系统硬件配置为：DELL PowerEdge R720 服务器，CPU 为 E5-2640*24 处理器 (2.5GHz)，内存 64G。图 3.3 显示的是在相同硬件平台和相同数据下不同算法的识别过程的耗时对比情况：

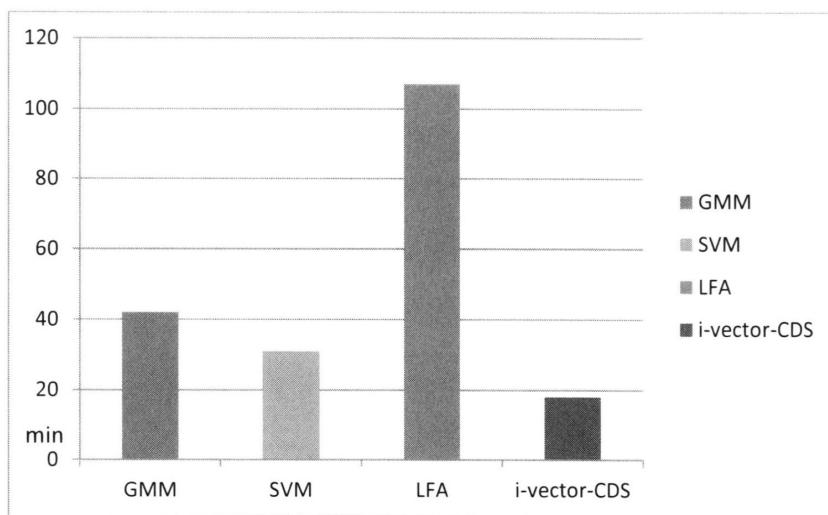


图 3.3 不同说话人识别算法耗时对比

通过对基于 GMM-UBM、GMM-SVM、LFA 和 i-vector 说话人识别系统性能和效率之间的横向对比，我们不难发现基于 i-vector 模型的说话人识别系统在识别效率上有明显的优势，在性能上虽然略逊于 LFA 算法，但是 LFA 算法在时间成本上消耗过大，在处理大规模数据时存在明显缺陷。

需要说明的另一点是，i-vector 算法不仅在计算复杂度上改善明显，同时算法的空间复杂度也表现优秀。因为个性特征经过投影降维后变为低维的矢量，无论在计算还是在存储上都远远小于高斯超矢量，提取出的 i-vector 文件大小在 4-6kb，这也为远程说话人识别的数据传输提供了极大的方便。

综上，基于 i-vector 模型的说话人识别系统以其高性能和高效率迅速成为说话人识别领域内的主流算法，也是说话人识别系统应用到大规模数据处理中的算法保证，不仅在理论上，有着重要的研究意义，在应用中也有着广泛的应用价值。

虽然 i-vector 算法取得了不错的效果，但也并非表现完美，在跨信道问题上依然有着不足之处，为解决这个问题，我们进行对信道补偿算法的研究。

3.4 信道补偿的 PLDA 算法

根据 i-vector 的原理可知，在全局差异空间中对说话人信息和信道信息不再区分，这就有可能存在提取出的 i-vector 不是最优的说话人特征，通过信道补偿算法有益于提高 i-vector 特征的准确性。

由于 i-vector 描述的是总变化因素，同时包含信道与说话者信息，因此可以直接在 i-vector 上执行信道补偿，而在 i-vector 上的信道补偿方法如类内协方差归一化（Within-class Covariance Normalization, WCCN）^[53]、线性区分性分析（Linear Discriminant Analysis, LDA）^[39] 以及概率线性区分性分析（Probabilistic Linear Discriminant Analysis, PLDA）^[40]，均有较好的效果，特别是 PLDA 改进尤其明显。图 3.4 是基于 PLDA 补偿算法的 i-vector 说话人识别系统流程图。

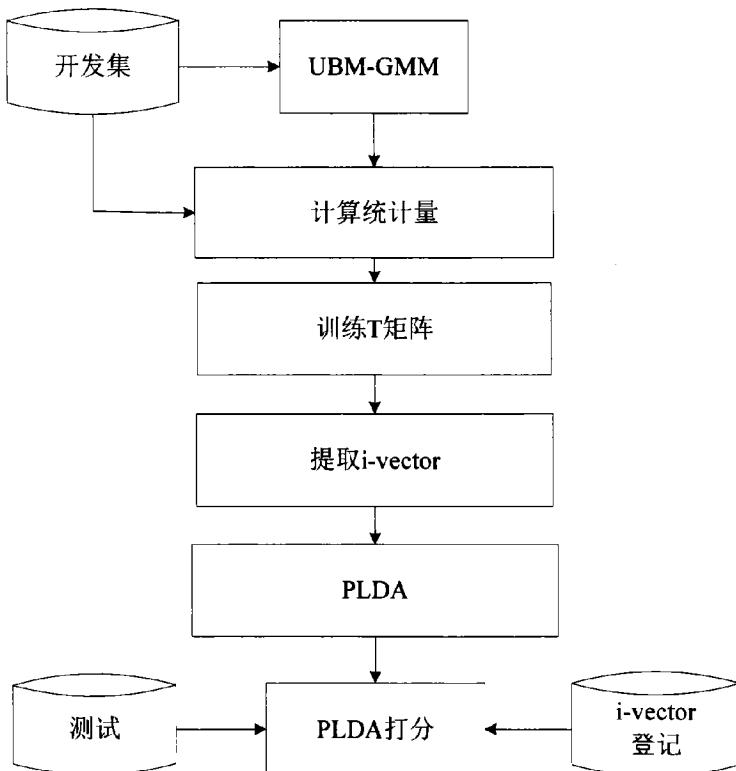


图 3.4 基于PLDA补偿算法系统流程图

PLDA 最初应用于人脸识别领域^[40], 后来被引用到说话人识别领域, 它与 JFA 类似, 可以看做是在 JFA 在单高斯下的一种特例^[57]。PLDA 对 i-vector 做进一步的因子分析, 具体表示如以下公式:

$$x_{ij} = \mu + \phi\beta_i + Gw_{ij} + \varepsilon_{ij} \quad (3-17)$$

其中 x_{ij} 是 i-vector, μ 是所有训练 i-vector 的均值, ϕ 是说话人空间矩阵, 用于刻画说话人特征, β_i 是对应的说话人因子, G 是信道空间矩阵, 用于描述信道特征, w_{ij} 是对应的信道因子, ε_{ij} 为残差因子。 β_i 和 w_{ij} 符合 $N(0, I)$ 分布。

在实际应用中, 往往把信道因子和残差因子进行合并处理, 公式 3.17 可以简化为:

$$x_{ij} = \mu + \phi\beta_i + \varepsilon_{ij} \quad (3-18)$$

其中 ε 满足高斯分布 $N(0, \Sigma)$, β 满足高斯分布 $N(0, I)$ 。这样对 PLDA 模型参数的估计就转化为根据训练数据对参数 ϕ 和 Σ 的估计。

3.4.1 PLDA 模型参数的训练

首先, 我们根据 x_{ij} 来定义一个统计量:

$$\begin{cases} \tilde{x}_i = \sum_{j=1}^{M_i} x_{ij} ; \text{(sufficient statistics for } i\text{th speaker)} \\ F_i = \frac{\tilde{x}_i}{M_i} ; \quad F_i \sim N(\phi\beta_i, \frac{\Sigma}{M_i}) \end{cases} \quad (3-19)$$

F_i 的分布用概率表示即 $P(F_i|\beta_i) = N(\phi\beta_i, \frac{\Sigma}{M_i})$ 。

根据贝叶斯法则, 有:

$$P(\beta_i|F_i) = \frac{P(F_i|\beta_i)P(\beta_i)}{P(F_i)} \quad (3-20)$$

即:

$$P(\beta_i|F_i) \propto P(F_i|\beta_i)P(\beta_i) \quad (3-21)$$

由于右边的两项都是高斯分布, 故左边的一项一定满足高斯分布。根据《Pattern Recognition and Machine Learning》中的公式, $P(\beta_i|F_i)$ 服从如下的高斯分布:

$$P(\beta_i|F_i) = N\{[I + \phi^T \Sigma^{-1} M_i \phi]^{-1} \phi^T \Sigma^{-1} M_i F_i, (I + \phi^T \Sigma^{-1} M_i \phi)^{-1}\} \quad (3-22)$$

估计 PLDA 模型的有效方法即为 EM 算法。该算法是一种迭代算法, 采用极大似然

估计求解含有隐变量的概率模型参数。在每一次迭代中，在 E 步骤先求出给定训练数据下隐变量的期望，然后在 M 步骤将这个期望最大化。通过迭代逐渐收敛，达到局部最优值。那么在 PLDA 模型参数的求解中：

E 步骤：

求在给定观测数据和当前参数下对未观测数据 β_i 的条件概率分布 $P(\beta_i|F_i)$ 的期望 $E(\beta_i|F_i)$ (在后文中简写为 $E(\beta_i)$)，即上式中高斯分布的均值。即：

$$E(\beta_i) = [I + \phi^T \Sigma^{-1} M_i \phi]^{-1} \phi^T \Sigma^{-1} M_i F_i \quad (3-23)$$

又由期望相关公式可以得到：

$$E(\beta_i \beta_i^T) = E(\beta_i)E(\beta_i^T) + [I + \phi^T \Sigma^{-1} M_i \phi]^{-1} \quad (3-24)$$

以上这两个期望会在后边的参数迭代求解中用到。

再进行 M 步骤的计算：

在这里，根据最大似然估计的原理，我们要最大化的是 $\prod_{i=1}^N \prod_{j=1}^{M_i} P(x_{ij}|\beta_i)$ ，取对数，有：

$$\begin{aligned} & \max \sum_{i=1}^N \sum_{j=1}^{M_i} \log[P(x_{ij}|\beta_i)P(\beta_i)] \\ &= \max \sum_{i=1}^N \sum_{j=1}^{M_i} [\log P(x_{ij}|\beta_i) + \log P(\beta_i)] \end{aligned} \quad (3-25)$$

由于 $P(x_{ij}|\beta_i)$ 和 $P(\beta_i)$ 都服从高斯分布，如下：

$$\begin{cases} P(x_{ij}|\beta_i) = N(\phi\beta_i, \Sigma) \\ P(\beta_i) = N(0, I) \end{cases} \quad (3-26)$$

这里给出高斯分布的概率密度函数的一般形式：

若 $P(x)$ 满足均值为 μ ，协方差为 Σ ，均值高斯维度为 K 的高斯分布，则 $P(x)$ 的概率密度为：

$$P(x) = \frac{1}{\sqrt{(2\pi)^K |\Sigma|}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right) \quad (3-27)$$

将 $P(x_{ij}|\beta_i)$ 和 $P(\beta_i)$ 各自的高斯分布的概率密度函数带入，稍加整理，可以得到：

$$\begin{aligned}
&= \max \sum_{i=1}^N \sum_{j=1}^{M_i} [-K \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \beta_i^T \beta_i - \frac{1}{2} (x_{ij} - \phi \beta_i)^T \Sigma^{-1} (x_{ij} - \phi \beta_i)] \\
&= \min \sum_{i=1}^N \sum_{j=1}^{M_i} [\frac{1}{2} \log |\Sigma| + \frac{1}{2} \beta_i^T \beta_i + \frac{1}{2} (x_{ij} - \phi \beta_i)^T \Sigma^{-1} (x_{ij} - \phi \beta_i)]
\end{aligned} \quad (3-28)$$

再经过一系列数学推导和矩阵理论知识，最终得到：

$$\begin{aligned}
&= \min \sum_{i=1}^N \sum_{j=1}^{M_i} [\frac{1}{2} \log |\Sigma| + \frac{1}{2} x_{ij}^T \Sigma^{-1} x_{ij} + \frac{1}{2} \text{tr}(\phi E(\beta_i \beta_i^T) \phi^T \Sigma^{-1}) + \frac{1}{2} E(\beta_i \beta_i^T) - \\
&x_{ij}^T \Sigma^{-1} \phi E(\beta_i)]
\end{aligned} \quad (3-29)$$

接下来，求似然函数的极值，只需要对 ϕ 和 Σ 求导。

将公式 3.23 和 3.24 中两个期望值代入，通过求导，最后得到：

$$\phi = (\sum_{i=1}^N M_i E(\beta_i)^T) (\sum_{i=1}^N M_i E(\beta_i \beta_i^T))^{-1} \quad (3-30)$$

$$\Sigma = \frac{\sum_{i=1}^N \sum_{j=1}^{M_i} [x_{ij} (x_{ij}^T - E(\beta_i)^T \phi^T)]}{\sum_{i=1}^N M_i} \quad (3-31)$$

参数可以随机初始化，然后通过 EM 算法的不断迭代（经验是迭代十次左右）直至收敛状态，便可以最终得到参数 ϕ 和 Σ 的值。从而得到 PLDA 模型参数的估计值，将训练得到的参数用于接下来的确认得分阶段。

3.4.2 基于 PLDA 模型的确认得分

在模型参数估计完毕后，对两个 i-vector x_1 和 x_2 进行确认打分，得分结果和由两个假设模型 H_0 和 H_1 的对数似然比值得到：

$$score = \log p(x_1, x_2 | H_0) - \log p(x_1, x_2 | H_1) \quad (3-32)$$

其中假设 H_0 表示两个 i-vector 属于同一个说话人， H_1 表示两个 i-vector 分属于不同的说话人。

$$\begin{aligned}
score &= \log N \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} u \\ u \end{bmatrix}, \begin{bmatrix} \Sigma + \Phi \Phi^T & \Phi \Phi^T \\ \Phi \Phi^T & \Sigma + \Phi \Phi^T \end{bmatrix} \right) - \\
&\log N \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} u \\ u \end{bmatrix}, \begin{bmatrix} \Sigma + \Phi \Phi^T & 0 \\ 0 & \Sigma + \Phi \Phi^T \end{bmatrix} \right)
\end{aligned} \quad (3-33)$$

其中 u 表示全局差异因子的均值，在测试时可以令 $u=0$ ，降低计算复杂度。

同时通过矩阵求逆，公式 3.33 可以化简为：

$$score = x_1^T Q x_1 + x_2^T Q x_2 + 2x_1^T P x_2 + K \quad (3-34)$$

其中： $Q = (\Sigma + \Phi \Phi^T)^{-1} - [\Sigma + \Phi \Phi^T - \Phi \Phi^T (\Sigma + \Phi \Phi^T)^{-1} \Phi \Phi^T]^{-1}$ ，

$$P = (\Sigma + \Phi \Phi^T)^{-1} \Phi \Phi^T [\Sigma + \Phi \Phi^T - \Phi \Phi^T (\Sigma + \Phi \Phi^T)^{-1} \Phi \Phi^T]^{-1}，$$

K 为常数项。

化简后，在测试阶段，系统复杂度降低，运算速度得到明显提升，根据 PLDA 模型的得分情况，可直接对识别结果进行判定。

3.4.3 PLDA 参数的 MAP 自适应

在研究中，经常存在研究的系统存在数据集样本不足的情况。假如有两个数据集 D_{in} 和 D_{out} ，其中 D_{in} 与待研究系统数据相匹配，但是数据量不足，在训练 PLDA 参数中不足以体现出 PLDA 的性能，而 D_{out} 含有丰富的数据样本，如果直接采用其训练的 PLDA 参数进行实验，就会存在数据不匹配的问题。

为此，本文提出了一种自适应方法，将数据集 D_{in} 和 D_{out} 的参数模型通过自适应更新得到新的参数模型^[61]，假设原始数据集 D_{in} 的 PLDA 模型参数为 $\{\phi_{in}, \Sigma_{in}\}$ ，数据集 D_{out} 的 PLDA 参数模型为 $\{\phi_{out}, \Sigma_{out}\}$ ，对两个数据集的模型参数采用线性加权的方式进行自适应，以到达对系统数据匹配的目的。基于最大后验概率（MAP）原理，参数自适应过程如下所示：

$$\phi_{adapt} = \alpha * \phi_{in} + (1 - \alpha) * \phi_{out} \quad (3-35)$$

$$\Sigma_{adapt} = \alpha * \Sigma_{in} + (1 - \alpha) * \Sigma_{out} \quad (3-36)$$

其中， α 为自适应因子，取值范围 0-1 之间。

自适应过程如图 3.5 所示：

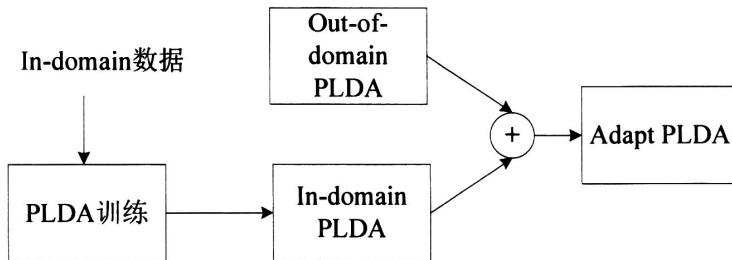


图 3.5 基于 PLDA 的领域自适应(domain adaptation)过程

3.4.4 实验结果与分析

为验证 PLDA 自适应的效果，本文采用数据集 D_{in} 为实网电话数据，包含 7,256 条实网语音数据，并通过训练得到其 PLDA 参数模型 $\{\phi_{in}, \Sigma_{in}\}$ ，数据集 D_{out} 为 NIST 数据集中 04,05,06 和 08 年的语音数据，共计 21,042 条。将这些语音进行训练，得到的 PLDA 参数模型为 $\{\phi_{out}, \Sigma_{out}\}$ ，测试数据采用实网电话数据，其中确认测试 1,512 次，冒充测试 1,871,856 次。

根据公式 (3-35) 和 (3-36)，在不同的自适应因子 α 下计算自适应后的参数模型 $\{\phi_{adapt}, \Sigma_{adapt}\}$ ，与参数模型 $\{\phi_{in}, \Sigma_{in}\}$ 和 $\{\phi_{out}, \Sigma_{out}\}$ 进行对比测试，结果如图 3.6, 3.7 所示：

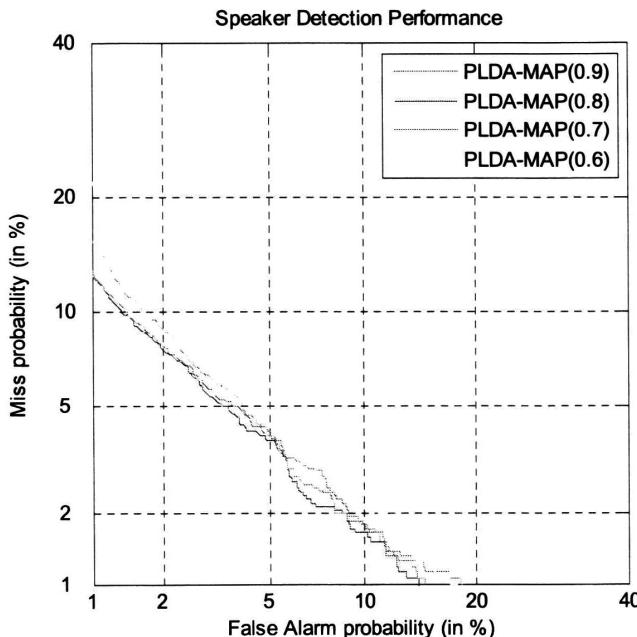


图 3.6 PLDA 的领域自适应不同自适应因子下性能对比曲线

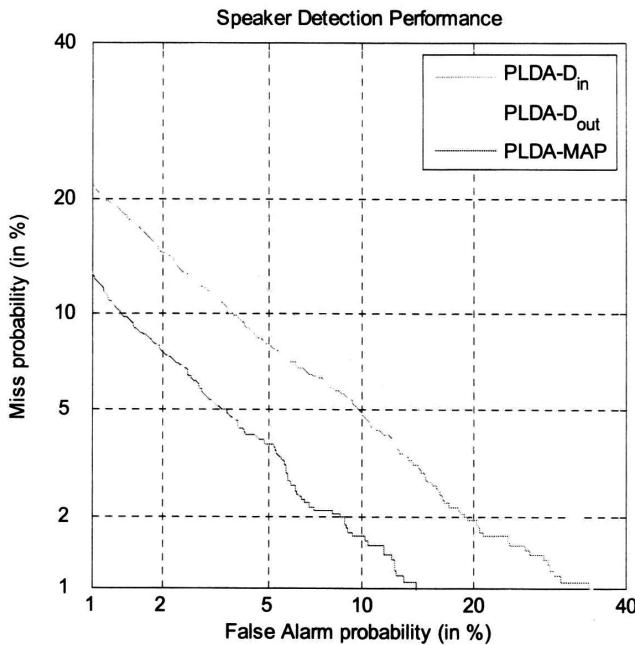


图 3.7 PLDA 的领域自适应性能曲线

对两个参数模型进行自适应，自适应因子取 0.9、0.8、0.7、0.6，考察自适应后系统的性能表现，由曲线可以看出，自适应因子在 0.8 处达到最好的性能，此时 PLDA-MAP 的 EER 为 4.1068%。由图 3.7 所示 PLDA-D_{in} 的 EER 为 6.6799%，这是由于训练数据有限，不能很好地表征说话人的语音特征。PLDA-D_{out} 的 EER 为 7.9365%，这是由于训练 PLDA 的数据与测试的数据库不匹配造成的。这是由于自适应后的模型融合了两个参数模型各自优势，所以可以具备更好的性能。

3.5 本章小结

本章详细介绍了基于 i-vector 模型的说话人识别系统，首先介绍了 i-vector 模型的基本思想，它不同于 JFA 将说话人因子和信道因子区分开来，而是将两者合并在同一个空间中，然后详细介绍了基于 i-vector 说话人识别系统的基本框架流程，从原理上推导了全局差异空间矩阵的训练过程，并对训练过程进行了理论分析，为接下来 i-vector 的提取做准备；在接下来的 i-vector 过程中，利用训练出的全局差异空间，将说话人语音投影到此空间上，得到一个低维的表征说话人个性特征矢量，简称 i-vector。而后对 i-vector 的打分过程进行了理论阐述和分析，在接下来的一节中，通过设计实验对比，验证了基

于 i-vector 说话人识别系统的高性能和高效性。

对于 i-vector 说话人识别系统对跨信道数据的鲁棒性问题，本章引入了 PLDA 的信道补偿算法，首先介绍了 PLDA 模型与 i-vector 结合的系统工作流程，然后从理论上分析了 PLDA 模型参数的具体训练过程，在 PLDA 模型参数估计的基础上，介绍了 PLDA 的测试打分方式，并在理论上降低了系统复杂度，提高系统的运算性能。

此页不缺內容

第四章 无监督聚类和说话人识别

4.1 引言

语音是人类最自然、最有效的交流方式，也是人类获取信息的重要途径。随着人类社会发展和信息化进程，特别是通讯技术、多媒体网络和互联网的迅速发展，语音服务和社交软件的盛行，人们在日常生活中对语音的需求和期待也越来越高，如何从海量的语音数据中迅速的获取我们想要的信息，如何准确有效的识别语音中的说话人信息，这在语音领域一直是个热点问题。

信息时代的到来，给人们生活带来便利的同时，也产生了海量的信息。对于语音方面来讲，互联网和通信网上每天产生的大量的语音数据，在这些数据中存在着大量的无效语音或无用信息，同时绝大多数都是无标签、无标记、无价值的“三无”数据。如何处理这些无标注数据，如何将这些数据恰当的运用在说话人识别系统中来提升系统的性能，这是本章重要研究的内容。

4.2 NIST2014 评测

NIST 评测^[58]分为说话人评测和语种评测两个分支，两个分支每年交替进行。NIST 评测自开始以来，一直是语音领域内的国际盛会，参与评测的研究者主要来源于国际知名机构有麻省理工的林肯实验室（MIT LL）、加州大学伯克利分校、剑桥研究中心（Cambridge）、捷克博诺技术大学（BUT）、国际商业公司（IBM）等；国内有清华大学，中科院、科大讯飞等。NIST 评测提供了一个统一的评测标准，使得各个研究机构、各类算法在相同的环境下得到客观公正的评价，也是领域内研究者之间相互交流，共同促进的重要机会。

2014 年的 NIST 说话人评测与往年有了较大不同，不再向参赛者提供原始语音文件，而是直接提供 i-vector 文件，要求参赛者在这些 i-vector 的基础上对系统的识别性能进行提升。没有原始语音文件，以往的说话人识别算法不再适用，研究者只能在基于 i-vector 的说话人识别系统中，引入机器学习的理论应对挑战。举办方人为的消除了不同研究者对语音前端处理的差异，只要求对系统的识别过程进行改进。这说明基于 i-vector 的说

话人识别技术已经成为说话人识别领域的主流算法，同时这也加强了说话人识别与机器学习的联系与融合，为研究者指明了今后主要的研究方向。

4.2.1 NIST2014 数据介绍

NIST2014 年，只向研究者提供的 i-vector 文件，是由 2004 年到 2012 年 SRE 语音文件的提取而来，全部来自于电话信道。每个 i-vector 是一个 600 维的向量。同时给出了每个 i-vector 对应语音的语音长度，所有的语音长度的平均时间为 39.58 秒。

提供的 i-vector 根据用途可以分为三种：模型集、测试集和开发集。

模型集，包含两个文件，一个为 i-vector 集合文件，另一个为 ID 对应文件。ID 对应文件中提供了目标说话人 ID 和目标说话人对应 i-vector 的 ID 之间的对应关系。目标说话人一共 1,306 个，每个目标说话人对应 5 条 i-vector。i-vector 文件中存放了所有目标说话人的 i-vector，一共 $1,306 \times 5 = 6,530$ 条。

测试集，提供了与模型集中无交集的 i-vector 集合，用于对系统识别性能进行测试。该数据集提供了 i-vector 的 ID，但对应目标说话人的 ID 未知。一共 9,634 条数据。

开发集，提供了与模型集、测试集无交集的 i-vector 集合，该数据集用于研究者对识别性能进行改进和研究。数据集提供了 i-vector 的 ID，但对应目标说话人的 ID 未知。同时要求研究者在研究和实验中，只能利用开发集中的无标注数据。一共 36,572 条。

ID 对应文件和 i-vector 集文件记录格式如图 4.1、4.2 所示：

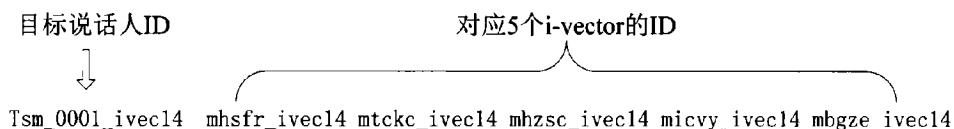


图 4.1 ID 对应文件记录格式

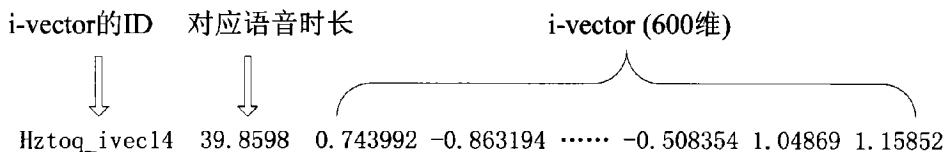


图 4.2 i-vector 集文件记录格式

根据要求，将模型集中的包含的所有目标说话人依次与测试集中的所有 i-vector 数据进行识别打分，然后将得到的打分结果保存到得分文件中。因此一共需要识别打分次数为： $1,306 * 9,634 = 12,582,004$ 次。

文章接下来的讨论和实验均是以 NIST2014 评测提供的数据为基础进行。

4.2.2 Whitening 规整

Whitening 规整^[59]是一种对矢量进行去相关处理的技术。它是对原有矢量进行空间转换，使转换后的矢量满足 $N(0, I)$ 分布。Whitening 规整步骤如下：

1、计算矢量空间的全局均值：

$$\bar{w} = \frac{1}{NJ} \sum_{s,h} w_{s,h} \quad (4-1)$$

2、对原矢量空间中心化以达到新矢量空间均值为 0 的目的。则新矢量空间的协方差矩阵为：

$$\Sigma = E((w_{s,h} - \bar{w})(w_{s,h} - \bar{w})^T) \quad (4-2)$$

3、如果矢量空间中存在矢量相关，那么协方差矩阵 Σ 将不是一个对角矩阵。因此需要对协方差矩阵进行对角化，以达到对矢量空间进行去相关的目的。为解决这个问题，可以借助于矩阵 Σ 的特征值和特征向量：

$$\Sigma \Phi = \Phi \Lambda \quad (4-3)$$

其中， Λ 为一个对角阵，由矩阵 Σ 的特征值组成； Φ 的列向量分别对应矩阵的特征向量。公式 4-3 改写为对角化形式为：

$$\Phi^T \Sigma \Phi = \Lambda \quad (4-4)$$

4、对一个单矢量进行对角化转换为：

$$\vec{w}_{s,h} = \Phi^T (w_{s,h} - \bar{w}) \quad (4-5)$$

此时 $\vec{w}_{s,h}$ 为去相关的矢量，其协方差 Σ' 为对角阵 Λ 。

5、在对角阵 Λ 中，其特征值可能相同，也可能不尽相同。让这些特征值都相同的过程就是对数据 whitening 的过程。每个特征向量的长度是由其对应的特征值决定的，所以未经过 whitening 规整的数据的协方差是一个椭圆分布^[59]，经过 whitening 规整后，特征向量的长度一致，符合一个圆分布：

$$\Lambda^{-1/2} \Lambda \Lambda^{-1/2} = \mathbf{I} \quad (4-6)$$

将公式 (4-4) 带入公式 (4-6) 中有:

$$\Lambda^{-1/2} \Phi^T \Sigma \Phi \Lambda^{-1/2} = \mathbf{I} \quad (4-7)$$

然后, 对去相关后的矢量 $w'_{s,h}$ 乘以上边的因子 $\Lambda^{-1/2}$, 就可得到 whitening 规整后的数据:

$$w''_{s,h} = \Lambda^{-1/2} w'_{s,h} = \Lambda^{-1/2} \Phi^T (w_{s,h} - u) \quad (4-8)$$

此时, $w''_{s,h}$ 的协方差矩阵是一个对角化矩阵, 同时也是一个单位矩阵。

$$E[\Lambda^{-1/2} \Phi^T (w_{s,h} - u)(w_{s,h} - u)^T \Phi \Lambda^{-1/2}] = \mathbf{I} \quad (4-9)$$

至此, $w''_{s,h}$ 就是矢量 $w_{s,h}$ 经过 whitening 规整后的矢量。

在 NIST2014 测评中, 利用开发集中的 i-vector 得到 whitening 规整矩阵, 然后对模型集、测试集和开发集中的所有 i-vector 进行 whitening 规整, 经过这样去相关后进行后面的研究与处理。

4.3 无标注数据的聚类实验与分析

通过上一节对 NIST2014 评测数据的介绍, 不难发现对开发集中大量的无标注数据的使用问题是这次评测的关键因素。但是在 PLDA 使用中, 但是需要大量的有标注数据 (一般要上万条, 说话人需要上千人) 进行训练参数。一方面是大量的数据无法投入使用, 一方面是需要大量的数据支撑算法, 因此对无标注数据的聚类和使用, 不仅对研究具有重要意义, 同时在应用中也具有很高的价值。

本节将先介绍无标注数据的基本情况, 在此基础上再探讨基于 SVM 算法和聚类 AHC 算法在无标注数据上的性能表现。

4.3.1 实验数据

实验原始数据来源于NIST2014评测官方提供的开发集, 包含i-vector记录36,572条, 每个i-vector为600维。为描述方便, 称原始开发集为开发集1.0、原始模型集为模型集1.0、原始测试集为测试集1.0。在实验前对所有的i-vector进行whitening规整, 得到开发集2.0、模型集2.0和测试集2.0。

同时, 在i-vector模型下, 无论语音长短, 均能够得到维度一致的i-vector。但是对于说话人特征的提取, 需要有足够长度的语音才能保证特征的准确性。因此根据开发集中

提供每条 i-vector 对应的语音长度，选择由长度在 20s 以上的语音提取出的 i-vector 进行实验，保证每条 i-vector 对说话人的特征信息有相对准确的刻画。经过挑选后，得到开发集 2.1，共有 24,057 条 i-vector。后续的实验以开发集 2.1 为数据来源。

4.3.2 基于 SVM 算法子系统

在 SVM 的这个子系统内，采用的是 libSVM 开源系统^[60]。libSVM 开源软件包主要用于解决分类、线性回归、分布估计等问题。在说话人识别中也是一种非常有效的工具。同时 libSVM 提供了不同的核函数^[60]，不同的核函数来源于不同的理论依据并会产生不同的实验结果，例如线性核（Linear Kernel）、多项式核（Polynomial kernel）、径向基核函数（Radial Basis Function，RBF）。鉴于 RBF 核对样本大小和维数高低均适用，同时可以将样本映射到更高维的空间内，最具有代表性、也最常用。因此在本实验采用了 RBF 核函数。

在实验开始前，首先对模型集和开发集进行分析与处理。因为开发集中的说话人与模型集中说话人无交集，因此在 SVM 分类中，特定说话人对应自身的 i-vector 作为正例，开发集中的无标注 i-vector 作为负例。根据上一节对模型集的介绍，每一个说话人对应的 i-vector 只有 5 条，也就是在对一个说话人的 i-vector 使用 SVM 分类时，有且仅有 5 个正例，如果把开发集 2.1 中所有的无标注数据作为负例，就会出现在一个待分类样本中，5 个正例对应 24,057 条负例的情况，这势必会造成 SVM 系统正负例极度失衡，同时也大大增加了系统的计算量。为解决这样的问题，可以考虑在开发集中随机选择 M 条 i-vector 作为负例。对于 M 的取值，既要保证有足够的负例矢量特征提升系统的性能和鲁棒性，又要考虑 SVM 分类系统的正负例之间的失衡问题。因此 M 的取值将会在下面的实验中，通过性能对比确定一个相对合适的值。

为了处理方便，根据官方给出 ID 对应关系文件，对模型集 2.0 中的 i-vector 进行处理，将每一个说话人对应的 5 条 i-vector 挑选出来，单独存放在以说话人 ID 命名的文件中，作为正例；然后将从开发集中随机挑选的 M 条 i-vector 同样添加在每一个说话人的文件中，作为负例。模型集中一共 1306 个说话人，因此会产生 1,306 个文件，每个文件中保存的是 (5+M) 条 600 维的 i-vector。

在上述准备工作完成后，开始搭建 SVM 分类系统，实验的具体过程为：

首先，将 i-vector 数据转换为 libSVM 软件包要求的格式。libSVM 工具包要求将输入数据格式为：Label 1: value 2: value ... 其中 Label 为类别标示。value 就是要训练的数

值，在这里就是 i-vector 中的各个特征值。因此将 1,306 个文件中的 5 个正例 i-vector 的 Label 设定为+1，M 个负例 i-vector 的 Label 设定为-1，i-vector 中每一维的特征数据前根据要求添加序号，直到第 600 维。

然后，利用 libSVM 中的 svmscale 工具，对原始样本进行缩放，缩放范围设定为[-1,1]。缩放的目的有两个：防止某个特征过大或过小，从而在训练中起的作用不平衡；为了使矢量单位化，加快计算速度。对 1,306 个样本文件进行缩放处理后，分别对各个样本文件数据进行更新。

其次，利用 svmtrain 实现对训练数据集的训练，获得 SVM 模型。在实验中，要对每个说话人的样本进行分类，经过训练后得到 1,306 个 SVM 模型文件。采用的 SVM 类型是 C-SVC，核函数类型为 RBF 核。

最后，根据训练获得的每个 SVM 模型，利用样本类别预测函数 svmpredict，对测试集中的所有 i-vectors 进行预测，在预测中输出测试集中全部 i-vector 基于每个 SVM 模型的概率估计预测值。

需要说明的是对于测试集中的 i-vector 数据，同样需要将其转换为 libSVM 软件包要求的格式，因为不清楚测试集中 i-vector 与模型集中说话人的对应关系，因此测试集中每个 i-vector 对应的 Label 值可以在+1 和-1 中任选其一，待系统预测后会给出正确的 Label 结果。

为确定较为合适的 M 值，对 M 分别取值 10、100、500、1,000、1,500、2,000、2,500、3,000、4,000、5,000 和 8,000 进行实验，实验结果如下图所示：

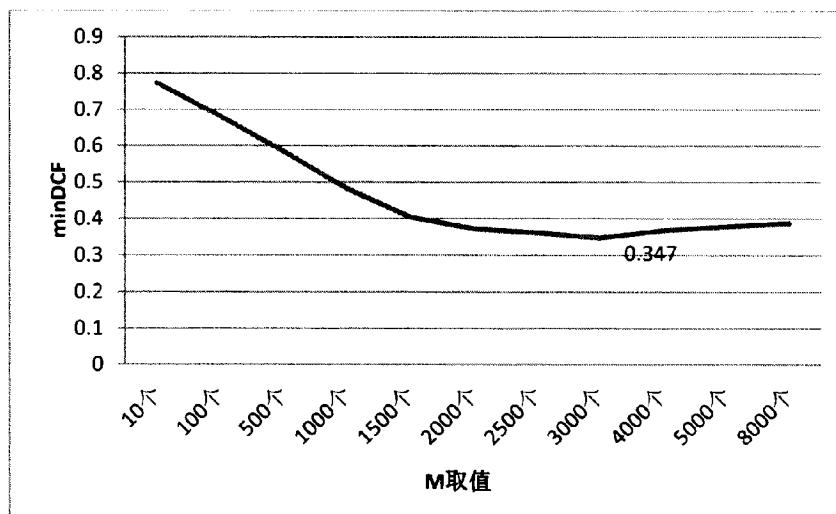


图4.3 M值对系统性能影响曲线

根据曲线变化趋势，在上述的取值集合中，M在3,000处时，系统对应的minDCF最小，即达到了最佳的性能。此时minDCF为0.347，对比使用Cos打分的基线系统的minDCF为0.386。由此可以得出，通过利用开发集中的无标注数据，使用SVM分类算法对系统性能提升了10.1%。

4.3.3 AHC 聚类算法

根据上一章中对PLDA算法的介绍，PLDA对系统的识别性能具有明显的改进。但是同时PLDA算法需要大量标注数据的支撑，因此我们考虑对开发集中的无标注数据进行聚类处理，把聚类结果中每一个聚类作为一个说话人，从而使用聚类结果训练PLDA的模型参数。

将无标注的i-vector自动的区分到不同的说话人特征类中，使用改进的经典AHC（Agglomerative Hierarchical Clustering）聚类算法^[61]进行聚类，AHC在语音领域被广泛应用在说话人聚类中。它是一种自下而上的聚类方法。

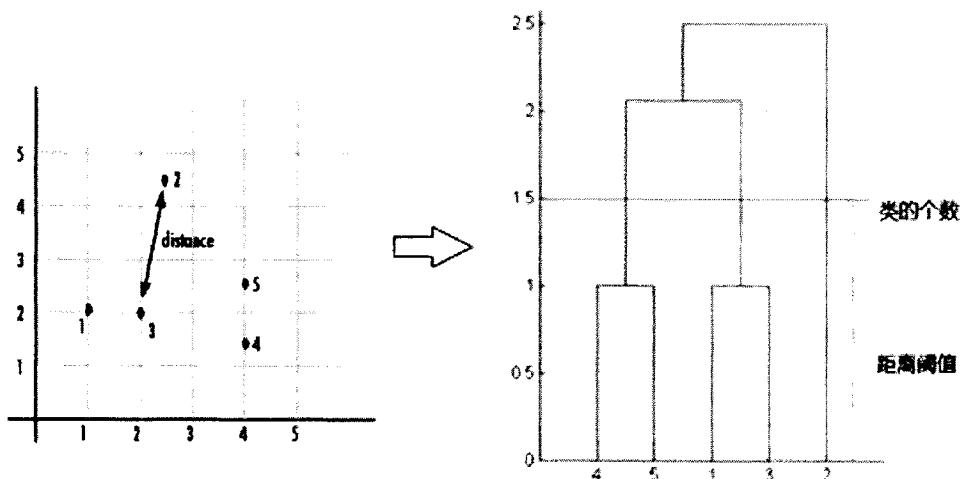


图4.4 AHC聚类原理

在经典AHC聚类算法中，对聚类数目的确定一直是聚类的难点问题，因为聚类的数目直接影响到聚类的效果。但现在需要用聚类处理的是大批量的无标注数据，对于聚类的数目无法提前确定。为解决这个问题，在聚类前不规定最终聚类的个数，而是根据样本之间的相似性估计值与一个阈值的比较结果，然后决定两个样本是否属于同一类。

为了更好地实施对i-vector的聚类，需要解决确定i-vector的性质和特性之间相似性估

计的统计量的问题。对于这个问题，i-vector之间的相似性估计一般使用余弦距离作为度量方法，它不仅方便和快捷，而且不需要对i-vector进行训练。因此，在聚类中使用向量之间的余弦距离作为判断标准。i-vector之间的相似性可以定义为： $\varphi(x, c) = \cos(x, c)$ 。

同时，在对于样本之间合并和类之间合并的问题，设计了三套不同的合并策略。第一种策略是根据样本间的距离进行排序，从小到大依次合并。第二种策略是在两个样本合并后，取两个样本的均值作为新的样本再与其他样本进行相似性估计的计算。第三种策略是一个样本点与类合并的判断依据是该样本点与类内所有样本之间距离的均值。同理，两个类之间也是根据类内所有样本点与另一个类中所有样本点之间的距离均值作为两个类之间的距离^[61]。

在聚类之前，首先需要设定阈值 τ_1 作为聚类合并的判定条件。同时规定数据输入和输出的文件格式：

$Data[L \times M]$ 是经过 whitening 规整后需要进行聚类添加标签的i-vector。其中 L 是 i-vector 数据量的个数，M 是 i-vector 的维数，这里 M 为 600。 τ_1 是判定是否将 i-vector x 吸收进类 c 中的阈值；

$\varphi(x, c) > \tau_1$ 将矢量 x 吸收进类 c 中

$\varphi(x, c) \leq \tau_1$ 将矢量 x 排除在类 c 外

输出数据：

$Labels[n \times h]$ 是聚类后添加标签结果，其中 n 是创建的类的个数， h 为每个类中含有 i-vector 的个数。每一行 i-vector 的 ID 代表它们属于同一类。

在上述工作完成后，讨论聚类的具体算法过程，三种策略的具体过程为：

策略一：

1. 每个 i-vector 设定为单独的一个类，并求两两之间的余弦距离；

For (i=1;i<=L)

 计算与其他所有类的距离：

 For(j=i+1;j<=L)

 计算 Cos(vector[i],vector[j])

 如果 CDS 大于 阈值 τ_1 ，将 i、j 和 CDS 信息存放到数组 sort[k++] 中。

2. 对数组 sort 根据 CDS 从大到小进行排序；

3. 根据排序后的结果进行合并；

While($m < k$)

 如果sort[k]对应的类i和类j，且类j中仅有自身一个i-vector，则将类j的label更新为i的label；

 如果类j中除了自身还有其他i-vector，将类j中所有i-vector的label更新为i的label；

4. 查找所有不同label的类，以及每个label中i-vector的个数，输出聚类后的结果

 如果label相同，n不变，h+1，同行输出；

 如果label不同，n+1，h=0，换行输出；

至此，聚类结束，得到由策略一算法下的聚类结果。

策略二：

1. 首先初始化各个i-vector的状态；

 将每个i-vector的label设定为自身ID，是否已被合并状态设定为false

2. 对i-vector进行聚类

For ($i = 1; i \leq L$)

 将vector[i]设定为一个类的中心；

 For ($j = 1; j \leq L$)

 如果vector[j]的合并状态为false

 计算Cos(vector[i], vector[j])

 如果CDS大于阈值 τ_1

 将vector[j]的label更新为vector[i]的label，vector[j]的合并状

态更新为true。若vector[j]同样也是一个类的中心，则将

vector[j]类中所有i-vector的label更新为vector[i]的label。

 将vector[i] = vector[i] + vector[j]，然后单位化；

 j再设定为1，重新循环。

3. 查找类的数目，并输出聚类结果

 如果vector为一个类的中心，同时合并状态为false

 n++，输出所有具有这个label的vector对应ID

至此，可以得到基于第二种策略的聚类结果。

策略三：

1. 每个i-vector设定为单独的一个类，并求两两之间的余弦距离；

For (i=1;i<=L)

计算与其他所有类的距离:

For(j=i+1;j<=L)

计算 $\text{Cos}(\text{vector}[i], \text{vector}[j])$ 并保存得到 $\text{Score}(\text{vector}[i], \text{vector}[j])$

2. 考察两个类之间的距离;

For (i=1;i<=num(Class1))

计算 Class1 中样本 $\text{vector}[i]$ 与类 Class2 中所有样本的距离:

For (j=1;j<=num(Class2))

$\text{Score} = \text{Score} + \text{Score}(\text{vector}[i], \text{vector}[j])$

求得分和的均值, 得到两个类之间的距离:

$\text{Score} = \text{Score} / (\text{num}(\text{Class1}) * \text{num}(\text{Class2}))$

3. 根据得分对 i-vector 进行聚类

如果 Score 距离大于阈值 τ_1

将类 Class1 和类 Class2 合并为一个类

5. 查找所有不同 label 的类, 以及每个 label 中 i-vector 的个数, 输出聚类后的结果

如果 label 相同, n 不变, h+1, 同行输出;

如果 label 不同, n+1, h=0, 换行输出;

至此, 可以得到基于第三种策略的聚类结果。

对于无标注数据的聚类, 同时也为了能够衡量不同算法之间聚类的效果, 假设了下面三个假设来确定聚类的有效性和性能:

1. 如果说话人 S 的类中绝大多数的 i-vector 都是属于说话人 S 的, 那么就认为对说话人 S 的分类是正确的;
2. 如果聚类后存在多于 1 个类都是属于说话人 S 的, 我们就认为含有最多 i-vector 的那个类是属于说话人 S 的类;
3. 为了防止多个类都属于说话人 S, 同时他们含有 i-vector 的数目也是相同的, 那么就确定一个类是属于说话人 S 的。

我们用聚类纯度 Q 的值来表征在一般的数据集上聚类的正确率^[61]:

$$Q = 100 \frac{M_{true}}{M} [\%] \quad (4-10)$$

其中, M_{true} 表示被正确聚类划分的 i-vector 的个数, M 表示聚类数据集中 i-vector 的

总个数。

同时，我们定义两个聚类效果会影响后续PLDA性能的错误率

Err_{ass} 是把不同说话人的i-vector分类到同一个类中的错误率；

Err_{sep} 是把同一个说话人的i-vector分类到不同类中的错误率；

我们定义一个类中有且仅有一个说话人的i-vector的类为一个“纯净”类，定义一个类中含有不同说话人的i-vector的类为“污染”的类，因此，把不同说话人的i-vector分类到同一个类中的错误率为：

$$Err_{ass} = 100 \frac{N_{con}}{N_{all}} [\%] \quad (4-11)$$

其中， N_{con} 是分类后被污染的类的个数， N_{all} 是分类的总个数。

把同一个说话人的i-vector分类到不同类中的错误率^[61]为：

$$Err_{sep} = 100 \frac{N_{bad-clear}}{N_{all}} [\%] \quad (4-12)$$

其中 $N_{bad-clear}$ 表示将本属同一类但是聚类后被拆分的类的个数，它就表现为将属于一个人的i-vector分类到不同的几个类中。

定义总的聚类错误率为：

$$Err_{sum} = Err_{sep} + Err_{ass} \quad (4-13)$$

为了考察聚类效果，同时对比三种策略下的聚类性能，实验选用NIST2014测评提供的模型集进行实验，因为模型集中6,530条i-vector的对应关系可以通过ID对应关系文件得知，因此通过对聚类正确率、聚类错误率等指标对聚类的性能进行科学评判。

通过策略一、策略二和策略三对模型集中的i-vector进行聚类实验，阈值设定为0.27^[61]，其对应的实验结果如表4-1所示：

表4-1 三种策略下聚类效果指标

| | Q | Err_{ass} | Err_{sep} | Err_{sum} |
|-----|---------------|---------------|--------------|---------------|
| 策略一 | 71.63% | 24.70% | 5.10% | 29.80% |
| 策略二 | 73.40% | 23.30% | 5.21% | 28.51% |
| 策略三 | 78.27% | 24.34% | 4.53% | 28.87% |

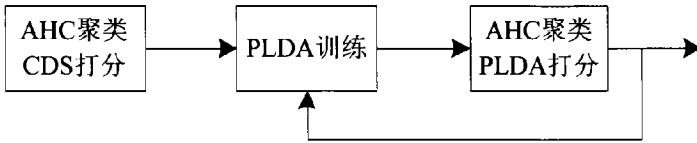
由三种策略进行聚类实验的效果来看，策略三和策略二在性能上比策略一稍有优势，根据理论分析，策略二中每个合并都利用新合并的i-vector对类的中心进行微调，从而保证在聚类过程中不会发生严重偏移的情况有关。而采用策略三是对类内中所有样本点之间的距离进行考察，对于样本点的吸收或者排斥在算法上相对“谨慎”与“民主”，这在一定程度上可以保证聚类的性能。但是三种策略整体性能都不理想，具有较大的错误率。对于后续的PLDA系统来说， Err_{ass} 和 Err_{sep} 会对参数的估计产生较大的影响，因此在性能上我们主要关注这两个错误率的指标。

通过对聚类结果文件的观察，发现在聚类结果中，会存在着有一个类中包含大量i-vector的情况，同时也有大量的类只含有一个i-vector。针对这样的情况，同时考虑到PLDA对数据的要求，我们对聚类结果进行处理。因为实验采用的是模型集的数据，因此每个类中至多有5个i-vector，根据这样的情况，我们将类中的i-vector数目大于10个或小于2个的类剔除，经过处理后 Err_{ass} 和 Err_{sep} 的指标得到大幅提升，此时 Err_{ass} 为11.2%， Err_{sep} 为0.347%。

另外根据对三种聚类策略的具体介绍，不难发现采用策略三在计算量上比策略一和策略二具有明显优势。策略一需要再计算所有i-vector的两两距离后再进行排序，然后根据合并策略进行聚类；策略二根据两个i-vector的距离对新样本合并后，需要对表征类的i-vector均值进行更新，在聚类过程中每次合并都有新的i-vector产生，这就使聚类过程中的计算量大大增加；而策略三只需在开始阶段计算所有i-vector的两两距离，在后面对两个类之间距离估计时，只需查找到对应的所有得分然后求得均值即可。策略三在聚类中降低了计算复杂度，节省了大量的时间成本，因此无论在实验中还是在应用中，采用得分均值策略的聚类算法都是一个恰当的选择。

4.3.4 AHC+PLDA 子系统

在上一节对 AHC 聚类的研究中，对无标注的 i-vector 数据之间的距离估计采用的是余弦距离算法。根据对 PLDA 的研究可知，利用 PLDA 对数据间距离进行估计比余弦距离具有更高的性能。在这样的基础上，将 AHC 聚类过程中的相似性估计过程，全部更换为 PLDA 打分过程。聚类基本流程为：首先利用余弦距离对无标注数据进行聚类，将聚类后的标注数据进行 PLDA 训练，然后利用训练出的模型，再对无标注数据进行重新聚类，不断迭代，最终达到最好的聚类性能^{[62][63]}。系统流程如图 4.5 所示：



4.5 AHC-PLDA 聚类流程图

需要说明的是，在采用 PLDA 打分作为无标注的 i-vector 数据之间的距离估计后，系统的计算量相对于余弦距离会大幅上升。第二种策略和第三种策略具有较强的聚类鲁棒性，但是第三种策略相对第二种策略具有较小计算量，因此采用第三种策略的聚类算法能够保证系统聚类的高效性。

待聚类的数据为开发集 2.1，其中包括 i-vector 共 24,057 条，且每条 i-vector 对应的语音均在 20s 以上，以保证对说话人特征的准确刻画。将所有无标注数据按照聚类要求格式整理后，进行聚类实验。

经过聚类后，每一个聚类可以认为是一个说话人，开发集 2.1 中的无标注数据通过 AHC 聚类后添加上了对应的说话人信息，此时是采用 CDS 进行打分，根据经验相似性阈值设定为 0.2，对应开发集更新为 3.0。开发集 3.0 中 24,057 条 i-vector 被划分为 7,573 个类，因为需要利用这些聚类结果进行 PLDA 模型参数的训练，而在 PLDA 训练过程中，根据经验，每个说话人对应的语音条数越多，PLDA 识别性能越好，系统鲁棒性也对应增强。因此我们对聚类结果进行进一步的处理。我们在开发集 3.0 的基础上，挑选出每个类中含有的 i-vector 数目不少于 3 个同时又不多于 30 个的聚类结果，得到开发集 3.1。这样选择不仅可以保证 PLDA 对训练数据的要求，同时可以提高聚类的纯度以保证 PLDA 训练参数的准确性。

通过 AHC 聚类和对聚类结果的提纯处理，为原始开发集中的无标注数据添加上了标签信息，变成了有标注数据。开发集 3.1 中的 i-vector 已经完全满足 PLDA 对训练数据的要求，在接下来 PLDA 的训练中，以开发集 3.1 作为训练数据。

利用开发集 3.1 进行 PLDA 的训练，为对比利用 PLDA 的打分与 Cos 的性能提升情况，对训练出的 PLDA 模型参数分别使用：一、在这些参数的基础上，利用 PLDA 测试数据进行打分，得到 AHC-Cos 聚类算法的识别结果；二、利用这些参数，PLDA 对无标注数据开发集 2.1 中的 i-vector 进行相似性评估，最终得到利用 PLDA 进行打分的聚类结果，采用 PLDA 得分进行相似性判断，根据经验相似性阈值设定为 5，将此聚类后的数据集标记为开发集 3.2。在开发集 3.1 中，共包含 1,626 个类，13,243 条 i-vector。

再在监督数据开发集 3.2 的基础上，利用 PLDA 算法对测试数据进行打分，最终得到 AHC-PLDA 聚类算法的识别结果。

在 PLDA 的训练中，为了对比通过聚类和 PLDA 组合对系统性能的提升，以及无标注数据经过聚类后作为 PLDA 的训练数据与有标注数据作为 PLDA 训练数据的对此情况，我们根据 NIST 官方后期提供的开发集的 ID 对应文件，将所有开发集中的数据加上标签后作为 PLDA 训练数据训练另一组 PLDA 参数模型，作为最好的效果进行比较。

为实验方便，实验数据全部选用模型集中的 i-vector，其中每个说话人对应 5 条 i-vector 中，选择 3 条作为模型集，剩余 2 条作为测试集。如此，模型集共 $1,306 \times 3 = 3,918$ 条 i-vector，测试集中 $1,306 \times 2 = 2,612$ 条 i-vector。在实验过程中，对每个说话人的 3 条 i-vector 求均值并单位化，作为该说话人的特征 i-vector。因此实验确认测试 2,612 次，冒充测试 3,408,661 次。

基于上述实验数据配置和算法选择，首先在三种聚类策略下，选择 AHC-Cos 聚类算和 AHC-PLDA 聚类算法对系统的识别性能进行对比。基于两种聚类系统的性能对比情况如图 4.6 所示：

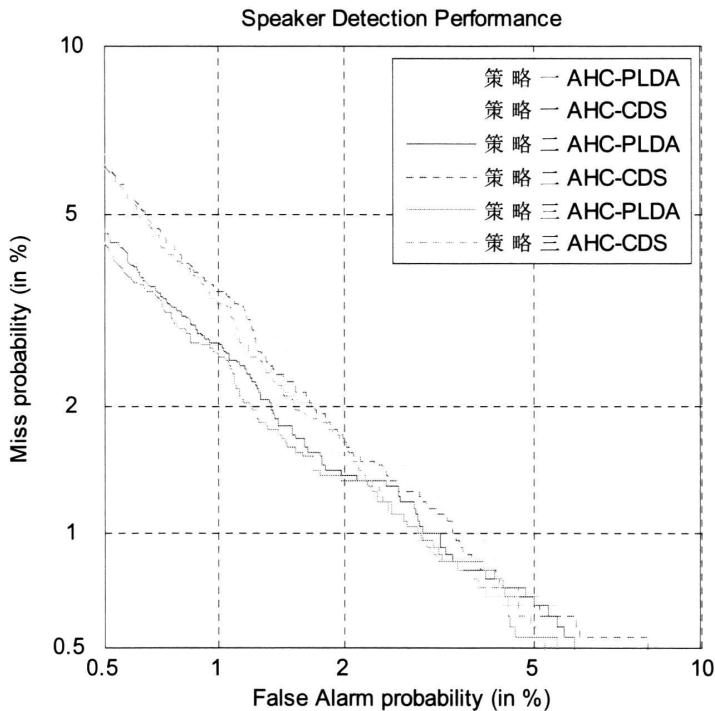


图 4.6 三种聚类策略下基于 CDS 和 PLDA 聚类实验结果曲线

根据实验曲线对比，以采用第三种聚类策略为例，基于 CDS 作为 i-vector 之间相似

性的判断标准得到的聚类结果,利用 PLDA 进行打分,系统的等错误率 EER 为 1.8111%;在利用 PLDA 作为相似性估计的聚类算法的聚类结果通过 PLDA 打分,得到的等错误率 EER 为 1.6127%。通过实验证明了利用 AHC-PLDA 的聚类算法在聚类的准确度上更具有优势。

同时,对于 AHC-PLDA 聚类算法在系统的迭代次数同样会对系统产生相应影响。随着 PLDA 的不断迭代,对聚类的准确度也会相应提升,从而系统的性能也不断提高。由于策略一和策略二算法复杂度较高,对迭代次数的实验仅在策略三的基础上进行。通过实验发现,随着迭代次数提高,系统性能会逐渐提升。但显而易见的是系统性能不会随着迭代次数提高而一直升高,随着聚类越来越准确,性能就会处于基本稳定状态。因此一味的追求提升迭代次数,达到提升系统性能的目的是以付出大量的计算为代价的。迭代次数对系统性能的影响如图 4.7 所示:

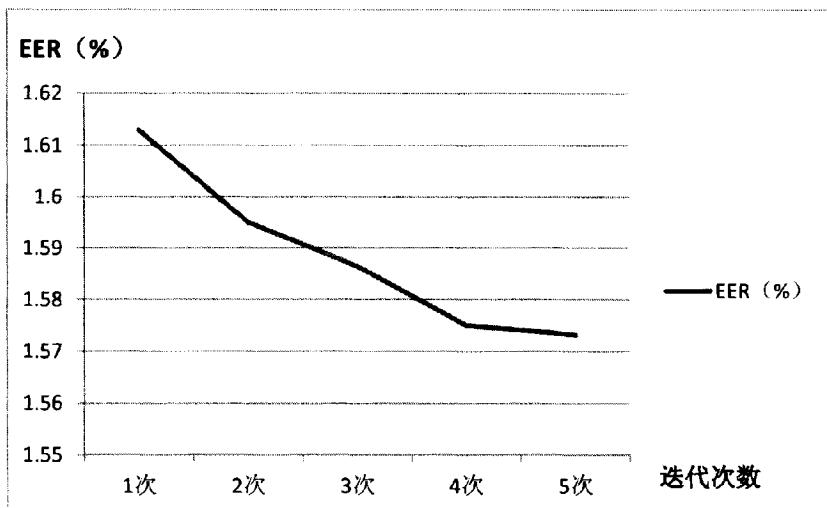


图 4.7 AHC-PLDA 聚类算法迭代次数对识别性能的影响

基线系统、AHC-PLDA 聚类+PLDA 打分、标注数据+PLDA 打分实验效果对比情况如图 4.8 所示:

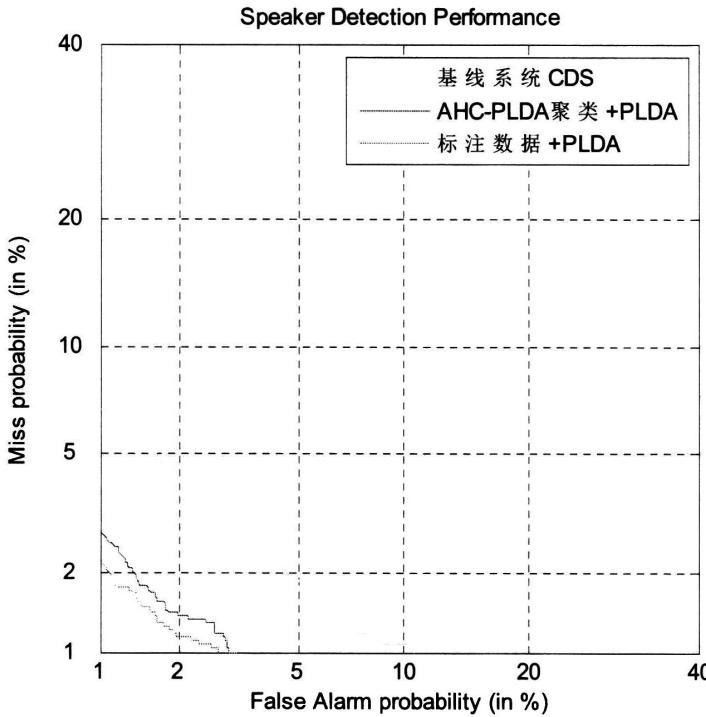


图 4.8 基线系统、AHC-PLDA 聚类+PLDA、标注数据+PLDA 对比实验结果

基线系统采用的是 CDS 直接打分，其等错误率 EER 为 3.0245%；基于 AHC-PLDA 聚类+PLDA 系统的等错误率 EER 为 1.5732%，开发集全标注数据+PLDA 系统的等错误率 EER 为 1.4931%。通过实验结果说明 AHC-PLDA 聚类算法和 PLDA 信道补偿算法不仅有效，同时明显提升了系统的识别性能。对比基线系统，AHC-PLDA 聚类算法+PLDA 算法对系统的识别性能具有明显提升。

更为重要的是，通过 AHC 无监督聚类算法和 PLDA 算法的结合，充分利用了无标注数据的价值，有效的架起了跨越无标注数据和有监督说话人识别之间鸿沟的桥梁。

4.4 系统得分融合

根据以往经验可知，对不同系统的得分结果在得分域上进行融合可以提高系统性能。融合过程类似于一个自适应的过程，可以根据基线系统的得分对新系统的得分进行微调和校准，以免系统的性能产生较大偏差。最常用的得分融合方法为线性融合^[62]，其基本思想是对新系统下得到的分数添加一个权重 α 因子，基线系统的权重因子为 $(1-\alpha)$ ，融合得分为两个系统的得分在权重因子作用下的得分之和。其计算公式为：

$$Score(i, j) = \alpha * Score_1(i, j) + (1 - \alpha) * Score_2(i, j) \quad (4-14)$$

其中 α 的取值范围为 0-1。同理，融合公式可推广到多个系统的得分融合，但需要保证所有权重因子之和为 1。

在本章中，为修正 PLDA 模型参数在某些特征点上可能产生的偏移，避免对系统的识别性能有一定影响，因此，采用基线系统的得分与基于 PLDA 模型下的得分进行融合，从而实现性能的进一步提升。

对得分进行融合前，首先需要确保不同系统的得分分布在同一个范围内，从而保证权重因子对得分之间的影响。基线系统采用余弦距离作为相似性判断的依据，其得分分布在-1-1 之间。而 PLDA 的得分结果范围跨度较大，且和余弦打分不在同一数量级，因此需要对 PLDA 的得分进行标准化规整，规整公式^[62]为：

$$Score(i, j) = \frac{(Score(i, j) - \min(Score))}{(\max(Score) - \min(Score))} \quad (4-15)$$

其中 $\min(Score)$ 为 PLDA 所有得分中的最小值， $\max(Score)$ 为所有得分中的最大值。经过规整后，PLDA 得分分布在 0-1 之间。对 PLDA 的得分进行规整，不仅可以对系统得分进行融合，而且在应用中对阈值的设定提供了极大的方便。

经过对 AHC-PLDA 子系统和基线系统的得分进行融合，对于融合因子 α 的确定，本文在 AHC-PLDA 聚类+PLDA 和标注数据+PLDA 两种系统下，分别与基线系统进行融合，对 α 分别取不同的值以确定一个合适的融合因子取值。

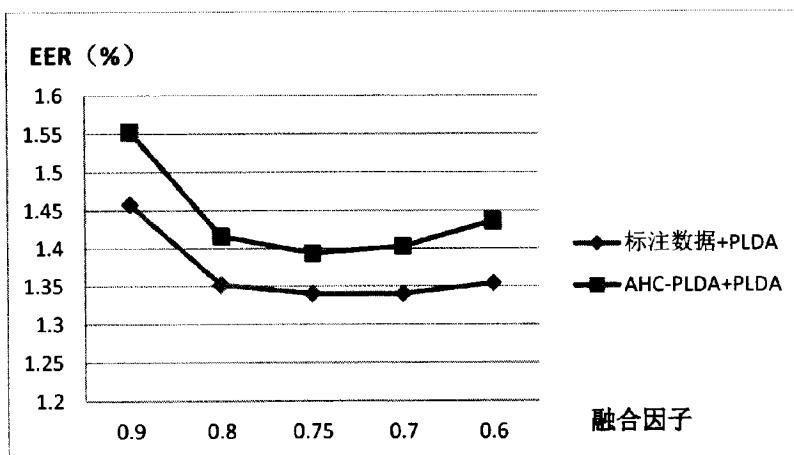


图 4.9 融合因子对系统融合性能的影响

根据实验结果，当 α 在取 0.75 时，两个系统的融合效果均较好。融合后的系统性

能对比如图 4.10 所示：

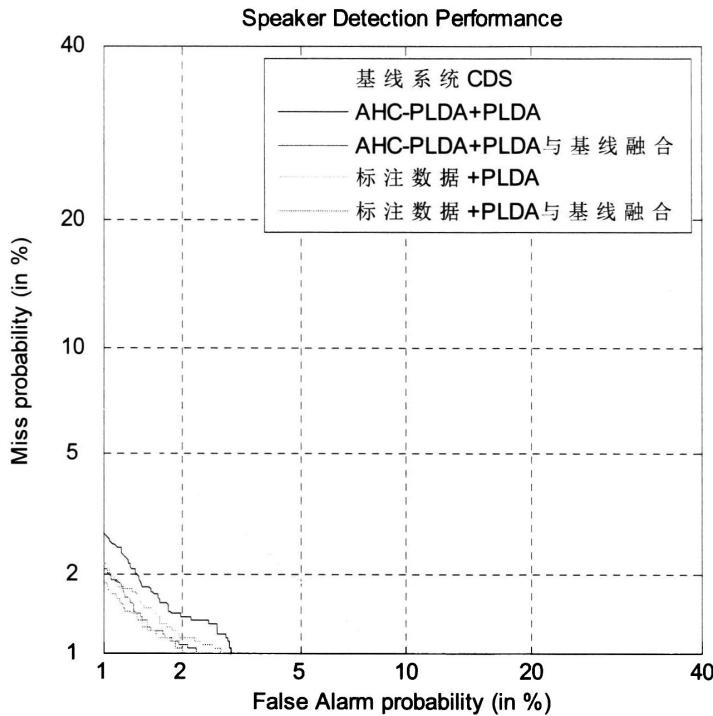


图 4.10 得分融合前后性能对比曲线

根据 DET 曲线可以看出, 经过对 PLDA 的得分乘以权重因子 0.75 与 CDS 得分乘以权重因子 0.25 之和得到新的得分文件, 通过这种融合后系统的识别性能得到了明显提升, AHC-PLDA 聚类算法+PLDA 系统与基线系统经过融合后, 其得分的等错误率 EER 为 1.3934%, 已经超过了在全标注数据+PLDA 的识别性能。全标注数据+PLDA 与基线系统经过融合后, 其等错误率 EER 为 1.3400%。相对于未融合前 AHC-PLDA+PLDA 组合的子系统, 识别性能提升了 24.3%。相对于用 CDS 的基线系统, 系统性能提升了 53.9%。

4.5 本章小结

本章主要介绍了在无标注数据下的无监督聚类和说话人识别问题。首先, 介绍了 NIST2014 说话人评测, 并对 NIST2014 年评测的主要内容以及官方提供的测评数据及其格式内容进行了详细说明。然后阐述了 whitening 规整的原理及作用, 该方法本质上就是对 i-vector 之间去相关处理, 以实现矢量之间的更好的差异性。对于评测的要求和目的, 针对评测提供开发集中的无标注数据使用问题, 成为了评测的关键因素。

因此，本章重点研究了基于无标注数据的使用方法。首先介绍了基于 SVM 算法的系统，文中使用无标注数据作为 SVM 分类中样本的负例，结合在评测中的实际应用对数据处理做了详细说明，同时在与基线系统的对比实验中，验证了在无标注数据下 SVM 算法对性能的提升作用。然后介绍了基于 AHC 聚类的基本理论，详细分析了该方法对无标注数据进行分类的原理，同时提出了针对 AHC 聚类的三种策略，然后针对这三种策略，详细阐述了三种策略的聚类过程；最后通过实验的方式对比三种策略的性能特点，为 PLDA 的训练过程提供科学有效的数据聚类工具。在接下来的一节中，利用 AHC-PLDA 算法聚类后的无标注数据，进行 PLDA 模型参数的训练，构建了 AHC-PLDA+PLDA 子系统，然后通过实验验证系统的性能，最终证明在无标注数据下，基于 AHC 聚类算法结合 PLDA 模型的打分系统相对于基线系统使得了系统性能提升了 39.1%。最后为了使算法的性能达到最好，本文在得分域对子系统得分进行了融合，通过系统融合，实现了系统性能的进一步提升。

本章针对无标注数据的分析和使用，在说话人识别领域不仅对算法的理论研究具有指导意义，同时面对大数据时代海量语音数据信息处理具有重要的应用价值。

此页不缺內容

第五章 说话人识别应用实例

5.1 项目意义

客户满意度直接影响着企业的品牌形象和销售收入，因此不少企业委托市场研究公司对其客户电话访问，进行满意度调查。如：汽车用户满意度是汽车厂商考核 4S 店的重要指标，也是年终返点的主要依据。因此不少 4S 店找人来冒充车主，接受市场研究公司的电话调查，虚假评价 4S 店的服务质量，严重影响了汽车厂商的客观考核。通过电话说话人识别技术，可对电访录音“听声辨人”，达到“去伪存真”的效果^[64]。

目前核查被替换联系电话的方法多数为数据库核查、核查电话取证等，但是通过人工进行核查造成人力成本高、核查周期长、经销商反核查能力提升等问题。说话人识别技术可以很好的解决这些问题，真正有效地将经销商找人冒充受访者的现实情况排查出来，巩固客户资料的真实性，以达赏罚分明的目标。

5.2 项目方案

根据对项目背景的介绍，系统的准确性和快速性成为项目的两个重要方面。同时由于项目语音来自于电话回访，电话信道的差别也带来了跨信道的问题。如何在保证识别高性能的前提下，提供高效的算法系统成为项目的关键。

根据项目的目的和具体要求，首先搭建基于说话人识别技术语音排查系统的基本结构框架：

首先，根据经销商提供被访问者的单方电话录音，对语音文件进行剔除静音和降噪等前端处理，以便获取相对纯净有效语音来提取更为精准的声学特征；

然后，对处理后的语音进行特征提取，根据说话人特征进行说话人建模，将说话人模型保存在说话人模型库中；

其次，将提取出的模型与模型库中的所有说话人模型进行一一对比，获得相似度得分后，与经验阈值比较，大于阈值判定为两个模型对应的语音来自同一个说话人，否则来自不同说话人。

最后，将系统判定属于同一个说话人的所有记录进行整理，得到最终说话人相似性

排查结果报告。

系统的模型提取过程和相似性排查过程如图 5.1 和图 5.2 所示：

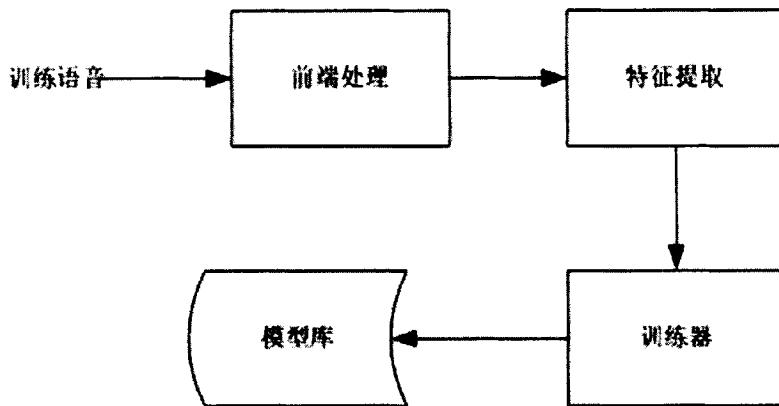


图 5.1 模型提取流程

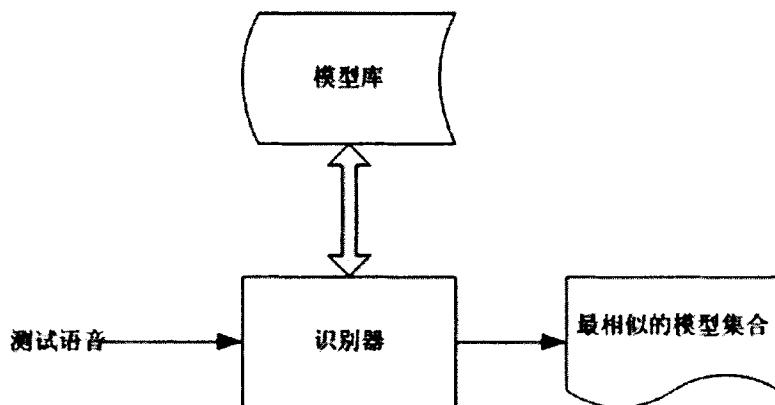


图 5.2 相似性排查流程

整体流程确定后，说话人识别系统的算法选择成为项目的关键因素。通过本文第二章对 GMM-UBM 系统和 GMM-SVM 系统的介绍，基于 SVM 算法的说话人识别系统在对语音的识别分类上具有明显的优势。但是由于电话语音信道的错综复杂，GMM-SVM 系统在信道鲁棒性方面表现不足，在跨信道识别中性能下降明显。LFA 算法虽然在抗噪和跨信道方面具有优势，但是 LFA 的计算复杂度在实际应用中远远不能满足项目对时间成本的控制。基于这样的条件，SVM 系统与 LFA 系统相结合成为一种解决方案^{[49][64]}。首先利用 SVM 系统对模型库中的相似组进行挑选，这个阶段认为是粗选的过程；然后 JFA 系统根据粗选的结果，对其结果进行二次挑选确认，找出最终的相似组，形成排查

报告。通过这种组合的方式，一方面可以在一定程度上提供排查准确度，同时很大程度上降低了单纯采用 JFA 系统的时间成本，加快了语音排查的效率。这种组合方式也就是项目最初的解决方案。

基于 i-vector 算法的说话人识别技术一经提出，以其优越的性能及系统的高效迅速吸引了广大研究者的目光。在本文第三章中，对 i-vector 的性能与传统算法的性能已经进行了对比，其性能已经得到了证明。另外关键的一点是，i-vector 的识别速度和效率被证明是目前算法中最高的，这在实际应用中具有很大的优势。同时在对大数据的处理中，空间复杂度和文件空间也是需要考虑的方面，相对于以往算法模型文件为几百 Kb，每个 i-vector 特征文件仅为 4k 左右，具有较大的存储优势。同时随着云计算的发展，说话人识别系统也将逐渐走向云平台的服务模式，在识别系统中，特征文件依靠客户端网络传送获得，在相同网络传输速度下，特征文件的大小直接关系到实际的用户体验。

项目的整体方案流程如图 5.3 所示：

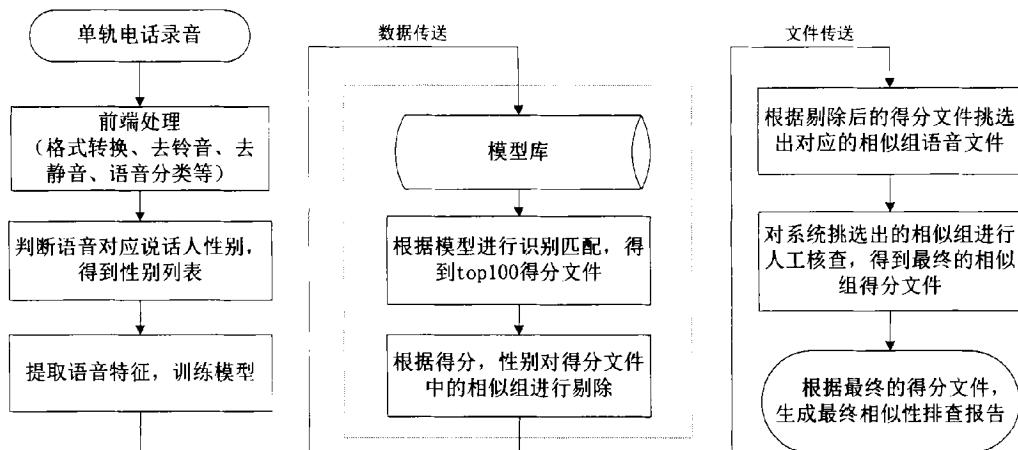


图 5.3 语音排查系统整体流程

整个系统流程可以在本地服务器端进行处理，最终提供排查报告。同时也可兼顾网络云平台基于说话人识别系统的语音排查服务的基本结构。图 5.3 中红色框图部分架设在网络服务器中，提供说话人识别的核心服务，其他部分均可架设在客户端，由于 i-vector 特征文件较小，对网络传输工作无太大压力。同时，特征文件的不可逆性保证了数据的存储和传输安全。

项目整个系统主要包括：语音预处理模块，声纹识别建模模块，识别结果确认核查模块，核查认证报告四个模块。

一、语音预处理模块

将用户满意度调查的语音文件进行建模前的预处理，其中包括录音文件去铃声、静音去除、格式转化、根据经销商提供的语音数据清单将录音文件分配到各自经销商文件夹下等过程。

系统在语音预处理模块的处理步骤如图 5.4 所示：

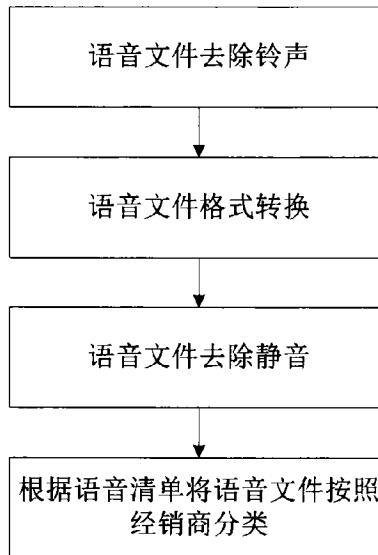


图 5.4 语音文件预处理模块

首先，系统对满意度调查语音进行去除铃声处理，排除语音文件中夹杂铃声，电话系统播报声对后续识别效果的影响。

其次，将去除铃声后的语音文件转换为系统处理的标准语音文件格式。

然后，将语音进行去除静音处理，减少无关因素对识别速度的影响。

最后，将上述处理过的语音文件按照经销商提供的语音清单进行分类，分别将语音文件分配到经销商对应的文件夹中。

二、声纹识别建模模块

所有的语音记录文件经过预处理后，在给定的 UBM 和全局差异矩阵的基础上进行 i-vector 的提取，将提取得到的 i-vector 存入对应的特征库中，为下一个模块中对 i-vector 进行对比识别做准备。同时，对语音文件中说话人的性别进行判断，得到语音文件对应说话人的性别记录。如图 5.5 所示：

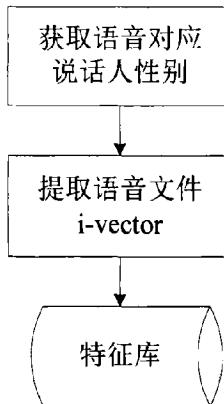


图 5.5 语音文件建模模块

三、识别结果确认核查模块

将待检测语音的 i-vector 文件与特征库中的所有 i-vector 进行打分确认，核查该条 i-vector 与基准语音的匹配得分是否大于阈值，如果大于阈值，同时两条 i-vector 对应语音的说话人性别一致，则该条语音记录保留；如果得分小于阈值或两者性别相异，则该条语音记录将会被剔除。最终经过确认核查后的相似组作为最终的相似组提交。

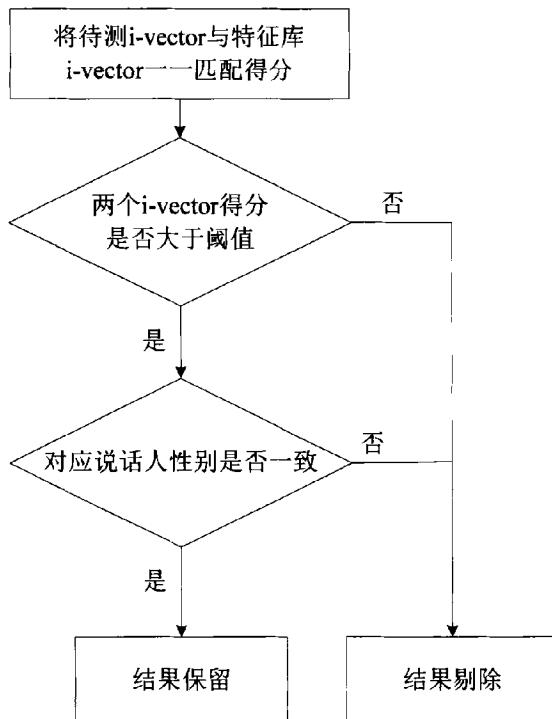


图 5.6 识别结果确认模块

四、核查认证报告模块

经过核查后的相似组作为最终的排查结果生成核查认证报告，在报告中会给出基准语音名，所属经销商，与基准语音相似的语音和所属经销商，相似程度得分等详细情况，如果不存在相似组，则说明不存在作弊的情况。

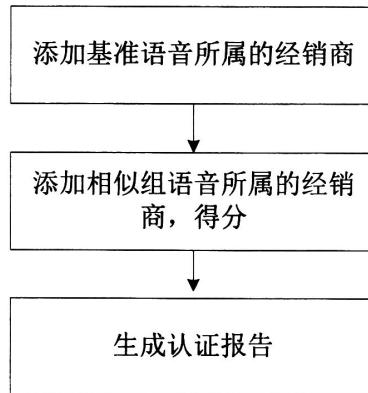


图 5.7 核查认证报告模块

为验证新系统的性能，本文对原有 GMM-SVM 和 LFA 系统进行性能的横向对比，实验测试数据采用实网电话数据，采用交叉测试方式。其中确认测试 1,512 次，冒充测试 1,871,856 次。系统性能测试对比曲线如图 5.8 所示。通过实验对比可以看出基于 i-vector 的说话人识别系统在性能对比上具有明显优势。

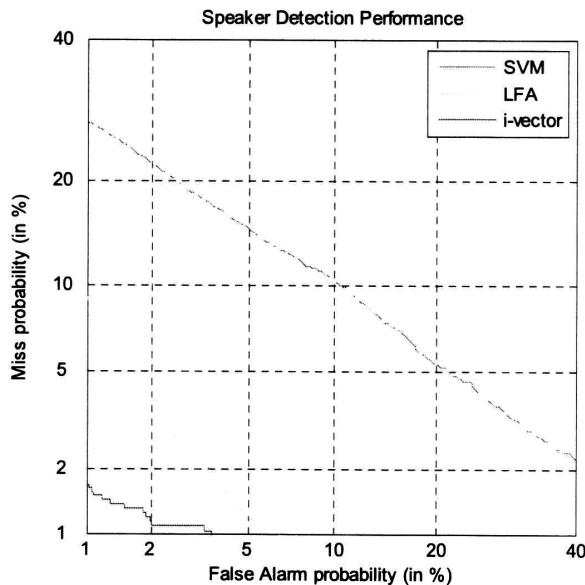


图 5.8 不同算法性能对比曲线

5.3 小结

说话人识别技术理论研究的迅速发展，带动了这种简单、高效的生物特征认证技术在应用领域的拓展。基于 i-vector 算法的识别系统，更加适用于大数据、高并发的应用。在互联网和移动客户端的产业布局中，说话人识别技术作为安全领域中的重要一员，势必将担负起更大的作用和责任。

此页不缺內容

第六章 全文总结及工作展望

本文主要是针对基于 i-vector 的说话人识别技术展开研究的。首先介绍了说话人识别技术中的经典的 GMM 和 SVM 识别算法，并在此基础引进了基于 i-vector 的说话人识别算法。然后本文对 i-vector 模型的结构和具体工作原理进行了阐述，对于信道鲁棒性问题，引入了 PLDA 信道补偿算法。其次结合 NIST2014 评测，对评测中的无标注数据进行分析和利用，用本文研究的 SVM 算法和 AHC 聚类与 PLDA 算法结合进行处理，根据处理实验结果验证对无标注数据的有效利用。最后，介绍了说话人识别技术结合语音排查实际应用，体现了说话人识别技术的应用价值。

将全文的工作内容总结如下：

1、说明了本文的研究背景及对现实的指导意义，介绍了国内外说话人识别技术的研究历史和现状，明确了本文的主要研究内容是基于 i-vector 模型的说话人识别，并在此基础上的 PLDA 信道补偿算法，以及在无标注数据下的说话人识别技术。

2、介绍了说话人识别的基本原理和结构，给出了在语音处理前端进行有效语音检测和特征提取的技术方法，在此基础上详细分析了经典的 GMM-UBM 说话人识别算法，并通过原理说明了算法存在的不足，在此基础上介绍了 GMM-SVM 模型，通过对 SVM 模型的介绍和工作原理，最后分析了这两种经典算法存在的问题和面对实际应用问题中的缺陷。

3、为了提高系统性能和效率，介绍了基于 i-vector 的识别算法，首先给出了该算法的基本原理，并在理论上详细推导了算法训练、识别和打分的公式，然后对算法的性能进行了分析，并通过实验与几个经典算法进行对比，验证了该算法的性能。最后针对基于 i-vector 说话人识别系统信道鲁棒性问题，研究了基于 PLDA 模型的信道补偿算法，通过对算法的工作原理的介绍和对 PLDA 模型参数估计及打分过程的理论推导，提供了 PLDA 补偿算法的科学依据。

4、重点描述 NIST2014 说话人评测情况，对评测给出的数据集和数据结构与格式做了详细说明。给出了对 i-vector 进行 whitening 规整的基本原理，并说明了规整的方法和对后续工作的作用。以评测内容中开发集包含的无标注数据为契机，提出了无标注数据在说话人识别中的有效利用问题。因此，重点研究了在无标注数据下的说话人识别技术。首先，介绍了 SVM 分类算法的基本方法，详细说明了对无标注数据的合理利用，即利

用无标注数据作为样本的负例，通过 libSVM 工具包对样本进行训练和识别，从而提升系统的识别性能，并用通过实验验证了该方法对无标注数据利用的有效性。然后，为了对无标注数据利用更加充分，提出了对无标注数据进行聚类的思路。在基于经典聚类算法 AHC 的基础上，介绍了对 AHC 聚类算法的三种改进策略，同时详细阐述了三种策略的具体工作思路和流程，并通过实验对比得出适合 PLDA 算法的聚类方法。最后，根据 AHC-PLDA 聚类算法对无标注数据的聚类结果，进行 PLDA 模型参数的训练，构建 AHC-PLDA 聚类+PLDA 子系统利用无标注数据对说话人识别的性能进行提升，并通过实验对算法效果进行验证。与基线系统对比，系统性能实现了 39.1% 的提升。经过对子系统得分进行融合后，系统的识别性能进一步提升，对比基线系统，系能性能提升 53.9%。

5、结合实际应用，介绍了说话人识别技术在实际项目中的应用。通过对项目和说话人识别解决方案的制定，分析了基于 i-vector 模型的说话人识别技术在应用场景中的独特优势和实用价值。

本文主要是对基于 i-vector 说话人识别系统进行了研究，针对信道补偿和对无标注数据的使用情况都取得了不错的性能提升。但是，还是有许多需要改进和进一步深入研究的方面，归纳为以下几点：

1、PLDA 信道补偿算法在实验中对系统的性能具有明显提升，由于 PLDA 模型参数需要大量说话人的一人多段语音进行训练，但在实际应用中，这样的数据集难以获得，由少量数据集训练出的 PLDA 参数对性能提升有限，将 PLDA 运用到实际应用中是后续需要开展的重点工作。

2、无标注数据聚类效果性能较低问题。本文讨论了基于 AHC 聚类的三种策略，虽然对部分无标注数据进行了有效利用，但是依然存在大量无标注数据没有得到充分运用。因此高性能的无监督聚类是一个需要深入研究的问题，也是一个很有实用意义的研究方向。

参 考 文 献

- [1] Prof. Jacob Benesty Dr., Prof. M. Mohan Sondhi Ph.D., Prof. Yiteng Arden Huang Dr.: Springer Handbook of Speech Processing [M],Cerman:Springer,2007
- [2] Joseph Keshet,Samy Bengio. Automatic Speech and Speaker Recognition : LArge Margin and Kernel Methods[M].Untied Kingdom, John Wiley & Sons Ltd,2009
- [3] Kevin Roebuck. Speech Recognition: High-impact emerging technology what you need to know:Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Pty Limited, 2011.
- [4] 韩纪庆,张磊,郑铁然.语音信号处理[M].北京:清华大学出版社,2004
- [5] Claussen, H.; Rosca, J.; Damper, R., "Mutual features for robust identification and verification", IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp.1849-1852, March 31 2008-April 4 2008
- [6] Savic, M.; Sorensen, J., "Phoneme based speaker verification," IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), 1992, vol.2, pp.165,168 vol.2, 23-26 Mar 1992
- [7] 张雪英,数字语音处理及 MATLAB 仿真[M].北京:电子工业出版社,2010.
- [8] Zbancioc, M.; Costin, M., "Using neural networks and LPCC to improve speech recognition.", International Symposium on Signals, Circuits and Systems(SCS) ,pp.445,448 vol.2, 2003
- [9] Yucesoy, E.; Nabiiev, V.V., "Comparison of MFCC, LPCC and PLP features for the determination of a speaker's gender," Signal Processing and Communications Applications Conference (SIU), 2014 22nd , pp.321,324, 23-25 April 2014
- [10] Xinxing Jing; Jinlong Ma; Jing Zhao; Haiyan Yang, "Speaker recognition based on principal component analysis of LPCC and MFCC,"IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 403,408, 5-8 Aug. 2014
- [11] S.Pruzansky. Pattern-matching procedure for sutomatic talker recognition. Journal of the Acoustical Society of America. 1963, 35(3):354-358.
- [12] JE.Luck. Automatic speaker verification using cepstral measurements. Journal of the Acoustical Society of America. 1969, 46(4B): 1026-1032.
- [13] F. Soong and A. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Transactions on Acoustics, Speech Signal Processing. 1988,36(6); 871-879.
- [14] D. Burton. Text-dependent speaker verification using vector quantization source coding. IEEE Transactions on Acoustics, Speech Signal Processing. 1987, 35(2): 133-143.
- [15] F.K. Soong,A.E. Rosenberg, B.-H. Juang and L.R. Rabiner. A vector quantization approach to speaker recognition. AT & T Technical Journal. 1987, 66: 14-26
- [16] S. Furui. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech Signal Processing. 1981, 29(2): 254-272
- [17] Yuan Liang, Xianglong Liu, Yihua Lou, Baosong Shan. An improved noise-robust voice activity detector based on hidden semi—Markov models. Pattern Recognition Letters, 2011:1044-1053.

- [18] Veisi H, Sameti H. Hidden-Markov-Model-Based voice activity detection with high speech detection rate for speech enhancement. *IET Signal Processing*. 20(4),2012:1145-1157.
- [19] D. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*. 1995, 17: 91-108.
- [20] R Kenny et al. Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech and Language processin*. 2007,15(4): 1448-1460.
- [21] W.Campbell et al. Phonetic speaker recognition with support vector machines. *Advances in Neural information Processing Systems*. 2004, MIT Press, Cambridge,MA.
- [22] K.Farrell, R. mammone and K. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on Speech Audio Processing*. 1994,2(1):194-205
- [23] D. Reynolds, T. Quatieri and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*. 2000,10(1): 19-41..
- [24] W.Campbell et al. Support vector machines for speaker and language recognition. *Computer Speech and Language*. 2006, 20(2-3): 210-229
- [25] L.Ferrer et al. Parameterization of prosodic feature distributions for SVM modeling in speaker recognition. *International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*. 2007, 233-236.
- [26] W.Campbell et al. Phonetic speaker recognition with support vector machines. *Advances in Neural information Processing Systems*. 2004, MIT Press, Cambridge,MA.
- [27] A. Hmich, A. Badri and A. Sahel. Automatic speaker identification by using the neural network. *Multimedia Computing and Systems (ICMCS) 2011*,1-5.
- [28] R.V. Pawar, P.P. Kajave and S.N. Mali. Speaker identification using neural networks. *World Academy of Science, Engineering and Technology*. 2005,12: 31-35.
- [29] F.WebersL.Manganaro,B.Peskin5and E.ShibergjUsing prosodic and lexical information for speaker identification,in Proc.of the International Conference on Acoustics, Speech, and Signal processing(ICASSP) 2002
- [30] D. Reynolds, The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus. In Proc. ICASSP 1996,pp. 113-116.
- [31] D. Reynolds, Comparison of background normalization methods for text-independent speaker verification. In Proc. Of the European Conference on Speech Communication and Technology (EUROSPEECH), 1997.
- [32] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, 2000, Score normalization for text-independent speaker verification systems. *Digital Signal Process*. 10(1-3), 42-54.
- [33] Wei Wu,Thomas Fang Zheng, and Mingxing Xu,Cohort-based speaker model synthesis for channel robust speaker recognitionJCASSP 06,May 14-19,2006,Toulouse,France,pp.I-893 896
- [34] R.Teunen,B-Shahshahani, and L,Heck,A model-based transformational approach to robust speaker recognition,in Proc.of the International Conference on Speech and Language Processing (IC-SLP),(Beijing),2000.

- [35] D. Reynolds, 2003, Channel robust speaker verification via feature mapping. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing(ICASSP 2003), Vol. 2, Hong Kong, China, April 2003, pp.53-56.
- [36] P. Kenny, 2006, Joint factor analysis of speaker and session variability: theory and algorithms. Technical Report CRIM-06/08-14.
- [37] A. Solomonoff, W. Campbell, and I. Boardman' Advances in channel compensation for SVM speaker recognition. In Proc. ICASSP 2005, pp.692-632.
- [38] Fauve B G B, Matrouf D, Schefifer N, et al. State-of-the-art performance in text-independent speaker verification through open-source software[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2007, 15(7): 1960-1968.
- [39] McLaren M, Leeuwen D A V. Sourcenormalised and weighted lda for robust speaker recognition using i-vectors.In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Prague, Czech Republic:IEEE, 2011. 5456–5459
- [40] Simon J D P, James H E. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007. 1–8
- [41] Dehak N, Karam Z, Reynolds D, Dehak R, Campbell W, Glass J. A channel-blind system for speaker verification. In:Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Prague, Czech Republic: IEEE, 2011. 4536–4539
- [42] <http://cc.ctiforum.com/jishu/hujiao/hujiaozhongxinjishu/yuyinshibieyufenxi/jishudongtai/345551.html>
- [43] <http://www.51callcenter.com/newsinfo/153/138110>
- [44] <http://htk.eng.cam.ac.uk/>
- [45] <http://www.signalprocessingociety.org/technical-committees/list/sl-tc/spl-nl/2013-05/ALIZE/>
- [46] 唐永锋,霍春宝.噪声环境下语音信号端点检测算法的研究与改进[J].人工智能及识别技术,2007.
- [47] 赵力.语音信号处理[M]..北京:机械工业出版社.2004.
- [48] 陈淑珍,张晨光,刘怀林等.基于改进的语音参数提取的线性预测[J] .武汉大学学报, 2003, (1) :91~94
- [49] 李兰泽,声纹识别系统在满意度电话回访作弊排查中的应用[D],硕士论文.厦门:厦门大学,2014.
- [50] Furui S. Comparison of speaker recognition methods using static features and dynamic features[J]. IEEE Transaction on Acoustics, Speech, and Signal Processing. June 1981, 29(3):342-350.
- [51] Shinozaki, T.; Kawahara, T., "GMM and HMM training by aggregated EM algorithm with increased ensemble sizes for robust parameter estimation," Acoustics, IEEE International Conference on Speech and Signal Processing(ICASSP), pp.4405,4408, March 31 2008-April 4 2008
- [52] R Kenny et al. A study of inter-speaker variability in speaker verification. IEEE Transactions on Audio, Speech and Language Processing. 2008, 16(5): 980-988.
- [53] N.Dehak et al. Front-end factor analysis for speaker verification. IEEE Transactions on Audio,Speech and Language Processing. 2011,19(4): 788-798.
- [54] N. Dehak, “Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification,” Ph.D. dissertation, École de Technologie Supérieure,

Montreal, QC, Canada, 2009

- [55] Fazzoli, Emilio. "Intro to Hidden Markov Models the Baum-Welch Algorithm". Aeronautics and Astronautics, Massachusetts Institute of Technology. Retrieved 2 October 2013.
- [56] N.Dehak et al Support vector machines and joint factor analysis for speaker verification IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),2009
- [57] 徐云飞,杨海,周若华,颜永红,高斯 PLDA 在说话人确认中的应用及其联合估计[J] 自动化学报 2014.6 vol.40 No.6
- [58] <https://ivectorchallenge.nist.gov/>
- [59] 蒋晔,基于短语音和信道变化的说话人识别研究[D],博士论文.南京:南京理工大学,2012
- [60] <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [61] Daniel Garcia-Romero et al. "Unsupervised domain adaptation for i-vector speaker recognition". The Speaker and Language Recognition Workshop 16-19 June 2014, Joensuu, Finland
- [62] Wenbo Liu,Zhidong Yu, Ming Li "An Iterative Framework for Unsupervised Learning in the PLDA based Speaker Verification" in Rroc ISCSLP 2014
- [63] Sergey Novoselov,Timur Pekhovsky, Konstantin Simonchik."STC Speaker Recognition System for the NIST i-Vector Challenge" The Speaker and Language Recognition Workshop 16-19 June 2014, Joensuu, Finland
- [64] <http://www.talentedsoft.com/Products/6.html>

攻读硕士期间的科研成果

- [1] Qingyang Hong*, **Sheng Wang**, Zhijian Liu.A Robust Speaker-Adaptive and Text-Prompted Speaker Verification System[C]. CCBR2014. 11 Shen Yang. China
- [2] 吕伟辰, 洪青阳, 王胜, 梁大为.基于 Viterbi-GMM 的文本提示型说话人识别系统[C]. 第十二届全国人机语音通信学术会议 (NCMMSC2013) , 2013 年 8 月, 贵阳
- [3] 中国发明专利: “文本提示型声纹门禁系统”, 专利申请号: CN201310294975.5
- [4] 中国发明专利: “基于声纹识别技术的满意度调查作弊排查方法”, 专利申请号: CN201310754586.6

此页不缺內容

致 谢

花开花谢，潮涨潮落，时光匆匆，白驹过隙。三年的研究生生活行将结束，在近 20 年求学道路的尾巴上回望过往，当初孑然一人来到这个美丽的城市的场景恍如昨日。看到有新的学弟学妹来学校面试，听到身边师弟们一句句师兄的叫着，无时无刻不在提醒着是时候需要离开了。

看到他们，想到了自己三年前的今天在干什么呢？三年前的今天我在为开启研究生生活的事情而四处奔波，当初来学校面试，我坐在海韵园的石凳上心里默默问自己：这会是我未来三年待的地方吗？不敢肯定又不忍否定。最终天随人愿，我得以机会在厦门大学度过这三年难忘的学习和生活。在临近毕业之际，最想表达的就是感谢，感谢所有给予我指引、关心和帮助的老师、朋友、同学及亲人。

首先要感谢我的导师洪青阳副教授，感谢他对我的知遇之恩，感谢他在学术上对我的无私指导，感谢他在生活上对我的关心和帮助。他把我带进了研究生的生活，在科研上对我不厌其烦的指引，让我在说话人识别领域有了明显的进步。不仅提高了我的科研能力，更重要的是教会了我面对问题怎样去寻求解决思路和方法，面对失败和挫折要坚定战胜它的信心。在他的悉心指导下，我的科研生活得以顺利开展。同时在他的帮助下，我得以在实际项目中锻炼自己的工作能力和办事效率。在生活上，他也给了我最大的帮助和关心。洪老师渊博的学识、严谨的治学态度、锲而不舍的科研精神、认真负责的做事风格以及对学生认真负责的态度对我产生了深刻的影响，将一直督促我不断学习，受益终身。在此，向敬爱的洪老师表达最诚挚的谢意！

感谢课题组李琳老师，通过与李琳老师的交流，她对我的指导和建议，对我的科研生活提供了很大的帮助，对我的科研和学习提供了强有力的支持。她清晰的思路和中肯的建议不断为课题组奉献着新的思路。我在研究生期间的成长离不开她的帮助和指导，感谢她的谆谆教诲和无私帮助。

感谢实验室和课题组的诸位师兄师弟师妹对我的支持和帮助，感谢雷文钿、李兰泽等师兄在科研工作中为我解决困惑，给予我很多的指导和帮助，让我少走弯路，拨云见日。感谢课题组各位成员在学习和生活上对我的关爱和帮助。感谢陪伴我共同度过三年研究生时光的刘翔鹏、陈炜超和黄玲，感谢他们在这期间对我的各种关心和帮助。感谢我的师弟张君和万丽虹师妹对我科研工作的支持和帮助。感谢我的室友王振、王彬彬和

李亮亮在生活上对我的包容和照顾。感谢我身边的同学们，在我采集语音和测试中给予了我了大量的方便与帮助。同时还要感谢天聪公司的各位同事，对我的工作提供了有力的协助和关照。

我还要特别感谢我的家人和女友，他们对我自始至终的鼓励和支持，是我成长的不竭动力，感谢他们这么多年来无尽的关心、爱护与包容，感谢他们多年对我的倾注的关爱和照顾。

最后，由衷的感谢论文的评审专家与答辩委员会的各位老师。