大规模目标说话人检测
关键技术研究

（申请清华大学工学博士学位论文）

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研 究 生 ： 王　　刚

指 导 教 师 ： 郑　　方　研究员

二〇一一年四月

Research on Key Technologies of Large Scale Target Speaker Detection

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

by

Gang Wang

(Computer Science and Technology)

Dissertation Supervisor： Professor Fang Zheng

April, 2011

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

为提高大规模目标说话人检测的性能，本论文针对多说话人分割问题和快速辨认问题进行了研究，主要工作包括：

1．提出基于参考说话人模型的距离度量算法。为提高在分析窗宽较短情况下两窗语音之间距离度量的稳定性，本文提出了基于参考说话人模型的距离度量算法。该算法不需要将窗内语音训练成模型，避免了数据较少、模型不精确对距离度量的影响，利用两窗语音分别与参考说话人模型之间的差异性来度量两窗语音之间的距离。与传统的 DISTBIC 算法相比，在 NIST SRE 2002 新闻采访语音库 BNEWS 上漏检率相对下降 34.8%，电话交谈语音库 SWBD 上漏检率相对下降 7.5%。

2．提出基于音素识别和文本相关的说话人分割算法。考虑到在较短语音下文本相关说话人识别好于文本无关说话人识别的原因在于文本相关信息的充分利用，提出通过音素识别技术获取音素这样的相关文本信息，以此进行文本相关的说话人识别的思路。在 TIMIT 数据库上，与基于参考说话人模型的分割算法相比，漏检率相对下降 15.4%。

3．提出基于类纯度约束的说话人聚类算法。为减轻分割后单说话人语音段长度过短对后续的说话人识别性能的影响，提出基于类纯度约束的说话人聚类算法。该算法借助参考说话人信息计算语音段之间的距离，以类内离散度最小、类纯度最大为准则，降低了不同说话人的语音被聚到同一类内的可能性。在 NIST SRE 2006 数据库上，在语音段平均长度分别为 2 秒、5 秒和 8 秒的条件下，与传统的 HAC 算法比较，有效类语音的比例分别提高了 2.7%、3.8%和 4.6%，目标说话人检测的召回率分别提高了 7.6%、6.2%和 5.1%。

4．提出基于参考说话人和双层结构的说话人快速辨认算法。目标说话人越多，说话人辨认所需要的时间越长，因此，大规模目标说话人辨认任务中辨认速度是必须面对的、极其关键的问题。为此，本文提出利用参考说话人度量待辨认语音与目标说话人之间的相似程度，并进一步利用双层结构进行剪枝以提高辨认速度的算法。在基于 GMM-UBM 架构的说话人辨认系统中，与传统的 SMC 算法相比，运算时间降低了 29.3%而辨认正确率提高 1.4%。

关键词：大规模；说话人检测；说话人分割；说话人聚类；快速辨认

Abstract

This dissertation focuses on the research on speaker segmentation and efficient identification issues to improve the performance of large scale target speaker detection task. It includes:

1. Reference speaker model (RSM) based distance measure. To improve the stability of distance measure between short windows, a speaker segmentation algorithm is proposed based on RSM, in which no model training is needed eliminating the influence caused by data sparseness and model inaccuracy, and the difference information of distances between the two windows against the reference speaker models. Compared with the conventional DISTBIC algorithm，the proposed algorithm can achieve a relative miss detection rate (MDR) reduction of 34.8% on NIST SRE 2002 BNEWS database and 7.5% on NIST SRE 2002 SWBD database.

2. Phoneme recognition and text-dependent speaker recognition based speaker segmentation algorithm. Taking it into account that text-dependent speaker recognition is much better than text-independent for short speech because text-dependent information is fully made use of, an algorithm using phoneme recognition technology to obtain phoneme text-dependent information is proposed for speaker segmentation. The distance is calculated between the same phonemes in two windows using text-dependent speaker recognition. Compared with the speaker segmentation algorithm based on RSM, the proposed algorithm can achieve a relative MDR reduction of 15.4% on the TIMIT database.

3. Class purity criterion based speaker clustering algorithm. To alleviate the influence on the performance of the following speaker recognition due to the average short length of single-speaker speech after segmentation, a speaker-clustering algorithm based on class purity criterion is proposed, where RSM is used to calculate the distance between speech segments and the minimal within-class dispersion as well as the maximal class purity are taken as the criteria. It reduces the probability of the speech segments by different speakers

being clustered into one same class. On the NIST SRE 2006 database, compared with the conventional HAC algorithm, for speech segments with average lengths of 2 seconds, 5 seconds and 8 seconds, the proposed algorithm can increase the valid class speech length by 2.7%, 3.8% and 4.6%, respectively, in the meanwhile the target speaker detection recall rate can be increased by 7.6%, 6.2% and 5.1%, respectively.

4. Efficient speaker identification algorithm using RSM based a Two-Layer Structure. It is obvious that the bigger the number of target speakers there are, the more time will be consumed for speaker identification. Therefore, the speaker identification speed is a crucial issue in large scale speaker identification task. To solve it, an algorithm using RSM to measure the similarity between the identified speech and target speaker is proposed. The proposed algorithm further uses a Two-Layer structure to prune target speakers and hence to improve the speaker identification speed. For GMM-UBM based speaker identification systems, compared with the conventional SMC algorithm, the proposed algorithm can achieve a computational time reduction of 29.3% and an identification performance increase of only 1.4%.

Key words: Large Scale; Speaker Detection; Speaker Segmentation; Speaker Clustering; Efficient Speaker Identification

目 录

第1章　绪论

语音是人与人进行交流的重要媒介，是最自然、最方便、最有效的交流工具之一，也是人类获取信息的主要来源之一。随着信息技术的不断发展，利用信息技术自动识别语音的说话人身份的技术也随之不断发展，即说话人识别（Speaker Recognition）技术[1]。说话人识别技术有着非常广阔的应用前景：在公安司法领域中，它可以用来寻找、发现、锁定和确认目标；在银行金融等领域中，它可以作为身份核对的一种手段，如声纹电话银行；在日常生活中，它可以用作个人身份的确定，如声控门禁等。大规模目标说话人检测是说话人识别的应用之一，目的是解决说话人识别所面临的由多说话人语音和大规模目标说话人等因素引发的问题，提高说话人识别的性能。论文在前人已有工作的基础上，对上述问题分别进行了研究，并提出了自己的一些见解。

本章的内容安排如下：1.1 对大规模目标说话人检测技术的组成与发展做简要介绍；1.2 综述大规模目标说话人检测的研究现状并指出其重点和难点；1.3 介绍本文工作的研究思路和工作内容；1.4 介绍本文的组织结构。

1.1　大规模目标说话人检测技术概述

大规模目标说话人检测技术是一种说话人识别技术，说话人识别是根据语音中反映说话人生理和行为特征的语音参数，来识别语音发出者身份的技术。

说话人识别根据应用的范畴可分为说话人辨认（Speaker Identification）[2]和说话人确认（Speaker Verification）[2]两类。说话人辨认是判定待识别的语音属于 N 位参考说话者中的某一位，是一个多选一的问题；说话人确认是确定一段语音是否由所声明的说话人发出，答案有“是”（接受）或“否”（拒绝）两种，是一个二选一的问题。

说话人识别根据识别的内容可以分为文本无关（Text-independent）和文本相关（Text-dependent）两类[2]。文本无关不指定说话人发音的文本，模型建立相对困难，但使用方便且应用范围较宽；文本相关在训练时要求用户按照指定文本发音，精确地建立每位说话人的模型（例如基于词、音素或音节的模型），在识别时要求用户必须按指定文本发音。一般来说，文本相关的说话人识别的性能要好于文本无关的说话人识别，但是文本无关的说话人识别应用的灵活性要大大好于文本相关。

说话人识别根据待识别语音的说话人可以分为闭集（Close-set）识别和开集（Open-set）识别两类[2]。闭集识别，待识别语音的说话人均属于已知的目标说话人集合（目标说话人也称作集内说话人，不属于目标说话人集合的说话人称作假冒者或集外说话人）。开集识别，待识别语音的说话人可能为集外说话人，即不属于已知的目标说话人集合。显然，开集识别的难度要大于闭集识别。

大规模目标说话人检测，其目的是检测输入语音中是否包含目标说话人发出的语音，其输入语音中一般包含多于一位说话人的语音（多说话人语音），目标说话人的数量多（大规模）。一般来说，说话人识别中待识别语音中只含有一位说话人的语音（单说话人语音），为避免混淆，在本文中将待识别语音为单说话人语音的说话人识别称为单说话人识别。单说话人识别对多说话人语音的处理存在着较大的问题，因为很显然将多说话人语音直接用来进行单说话人识别在理论上是说不通的，无论多说话人语音与某一说话人匹配地如何完美也不能够说明这段语音是由这位说话人发出的，因为这段语音中还包含了其他说话人的语音不能代表单一某位说话人的特性。

大规模目标说话人检测可以分解成两个子任务：一是将多说话人语音转换成多段单说话人语音。这就需要检测多说话人语音中不同说话人说话的时间点，根据语音中说话人身份发生变化的时间点将多说话人语音分割成许多小段语音，这也是通常所说的说话人分割（Speaker Segmentation）[3]。因为分割之后的单说话人语音段的长度可能较短，会对单说话人识别的性能造成一定的影响，说话人分割之后一般还需进行说话人聚类（Speaker Clustering）[3]，说话人聚类是将说话人分割之后的语音段按照说话人的身份进行聚类，将属于同一说话人的语音段聚成一类。二是单说话人识别任务，在说话人检测中通常是指说话人辨认，即对说话人分割聚类之后得到多段单说话人语音分别进行说话人辨认，在辨认结果中回答多说话人语音中否有目标说话人发音，如果有目标说话人发音的话回答是哪些目标说话人的发音。图 1.1 是说话人检测的问题分解示意图。

大规模目标说话人检测系统的输入语音是多个说话人的任意文本的随意发音（多说话人语音、文本无关）、输入语音中既包含目标说话人的语音也包含假冒者的语音（开集）、目标说话人的数量多（大规模目标说话人）；没有输入语音的先验知识，如输入语音中的说话人数量、性别以及可能的目标说话人身份等；目的是检测输入语音中是否有目标说话人发出的语音，如果有的话是哪一个或哪几个目标说话人。因此，本论文所研究的系统是一个文本无关的大规模的开集的多说话人检测系统。

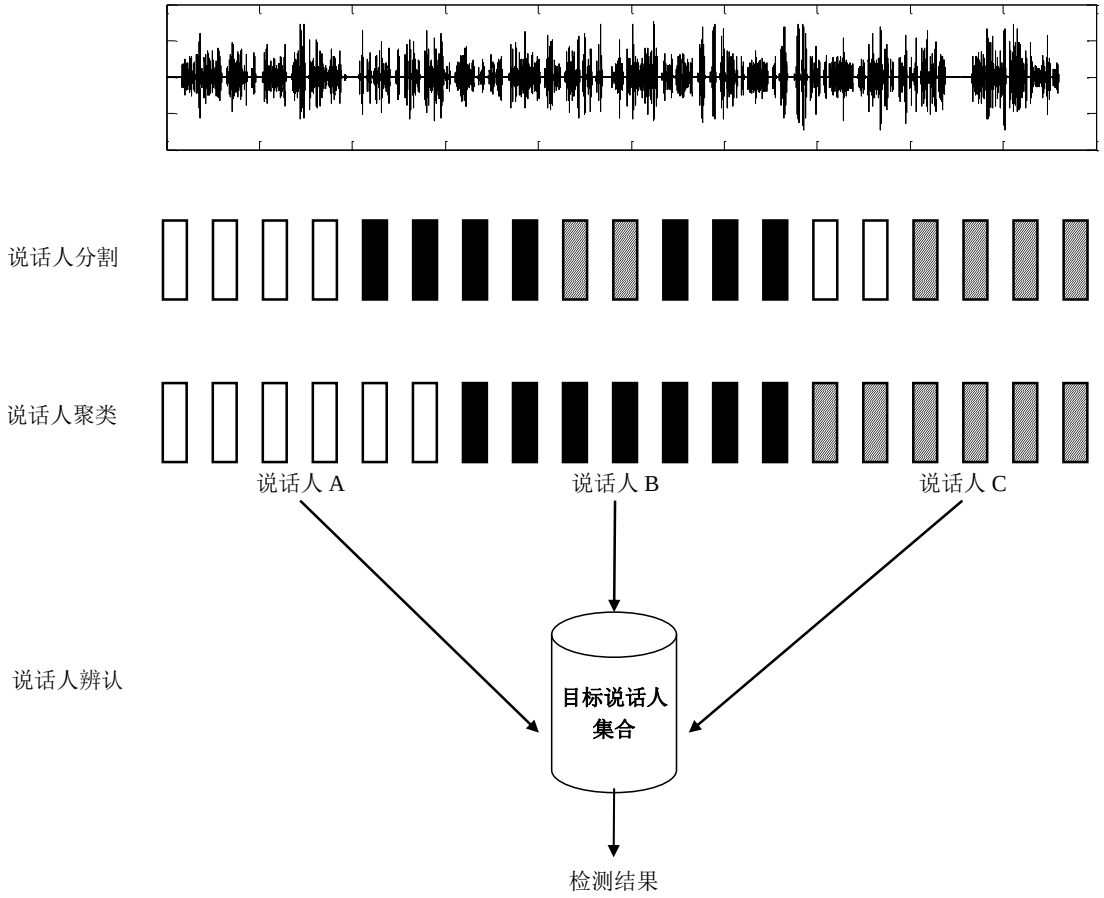


图 1.1 说话人检测问题分解示意图

当目标说话人集合过大时必然会造成辨认速度的大幅下降[4]，因此本文的研究任务可分解为两部分，一是提高将多说话人语音转换成多段单说话人语音的能力，即改善说话人分割聚类的性能；二是在保持说话人辨认的辨认准确率的前提下提高辨认的速度。

1.2　大规模目标说话人检测技术的研究现状

说话人识别的研究始于 20 世纪 30 年代，几十年来国内外许多著名大学、研究机构以及很多大公司的实验室都在开展说话人识别方面的相关研究，并取得了丰硕的研究成果，国外的有美国的 AT&T 实验室、麻省理工学院林肯实验室（Lincoln Laboratory）、加州大学伯克利分校的国际计算机科学研究院（International Computer Science Institute，ICSI）、IBM 的 Watson 研究中心等，加拿大的 CRIM（Centre de recherche informatique de Montréal，CRIM）实验室等，法国的 LIA（Laboratoire Informatique Avigonon，LIA）实验室、CLIPS（Communication

Langagiere et Interaction Personne Systeme，CLIPS）实验室和 LIMSI-CNRS 实验室等；国内的有中国科学院声学研究所、自动化研究所，北京大学，中国科技大学、科大讯飞语音实验室，北京邮电大学，北京理工大学，上海交通大学，浙江大学，南京大学，哈尔滨工业大学等。

近十多年来，输入语音为多说话人语音的说话人检测也逐渐开展起来。美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）[5]在 1999 年组织的说话人识别评测（Speaker Recognition Evaluation，SRE）[5]中就有了双人语音的说话人检测任务；NIST SRE 2002 和 2003 中组织了说话人分割聚类评测；NIST 在 2004 年以后的 RTE（Rich Transcription Evaluation，RTE）[6]当中，将说话人分割作为 Speaker Diarization 评测的一项子任务。一些说话人检测系统也逐渐出现，如法国的 ELISA 系统[7-9]、MultiStage[10,11]系统等。大规模目标说话人快速辨认的研究在近些年也取得了相当大的进展，研究者们提出了很多的快速辨认算法，取得了很好的加速效果[4]。

下面，本文将主要从说话人分割聚类和说话人快速辨认两个方面来介绍研究进展并给出分析。

1.2.1　说话人分割聚类的研究现状

1.2.1.1　说话人分割聚类的常用算法

总的来说，说话人分割聚类算法可分为三类：（1）基于距离度量的算法，该类算法主要是利用一些距离度量准则，利用预先设定好的阈值判断相邻的两段语音是否属于同一说话人。常用的度量准则有贝叶斯信息准则（Bayesian Information Criterion，BIC）[12-16]、一般化似然比（Generalized Likelihood Ratio，GLR）[17-20]、KL 距离（Kullback-Leibler Distance）[21-24]、交叉似然比（Cross Likelihood Ratio，CLR）[25,26]和权重欧式距离[27]等。（2）基于模型搜索的算法，该类算法需要已知目标说话人模型，或从多说话人语音中估计出可能的目标说话人模型，利用这些模型来搜索目标说话人的发音时刻，不断的迭代更新目标说话人的模型并对输入语音进行重搜索来完成说话人检测；常用的算法有美国 AT&T 实验室的基于高斯混合模型（Gaussian Mixture Model，GMM）的搜索算法[28]、法国 LIA 实验室的基于隐马尔科夫模型（Hidden Markov Model，HMM）的搜索算法等[29]。（3）距离度量和模型搜索相融合的算法，如法国的 ELISA 系统[7]，MultiStage 系统[10]等。目前，一些国内外研究机构均已经出现了不少具有实用价值的说话人分割聚类算法和系统，以下是其中一些的简介。

（1）法国 LIMSI-CNRS 的多阶段（MultiStage）[10]说话人分割系统，首先利

用语音活动检测（Speech Activity Detection, SAD）将语音分成语音和非语音两类，接下来使用基于 KL 距离的说话人分割算法进行初始分割，将分割得到的每个语音段训练成 GMM 模型[30]并利用 Viterbi 解码算法重新分割，然后使用 BIC 作为距离度量准则对分割后的语音段进行模型聚类，将属于同一类的语音段合并，并重估合并后的语音段的 GMM 模型，利用带有能量限制的 Viterbi 解码算法重新分割，最后按照说话人进行聚类。MultiStage 系统在 NIST RT 04F[6]和 ESTER[31]评测数据集上，对比单阶段 BIC 系统的说话人错误率[6]相对下降了 40%。

（2）美国 AT&T 实验室提出了一种基于 GMM 的说话人检测算法[28]。该算法首先在训练阶段利用已知的目标说话人语音训练目标说话人模型，检测时根据语音段在目标说话人模型和背景说话人模型上的似然分数差进行目标说话人检测。在 HUB4 新闻数据库[32]上，对于单一目标说话人检测，在语音质量很干净的情况下，漏检率大约是 7%，在语音质量不干净的情况下，漏检率大约是 27%；对于双目标说话人检测，漏检率大约是 63%。

（3）法国的 ELISA 系统[7]，该系统将基于 BIC 的 CLIPS 系统和基于 HMM 的 LIA 系统进行融合，融合策略有串行和并行两种，串行融合是将 CLIPS 系统的输出作为 LIA 系统的输入，并行融合是将 CLIPS 系统和 LIA 系统的结果首先进行融合，分割结果一致的部分保留不变，对于分割结果不同的语音段，采用任一系统进行重新分割。在 NIST SRE 2002，2003 和 2004 评测中，ELISA 系统分别取得了会议交谈语音库和电话对话语音库上的最优性能[7]，最优系统性能[8]和最优说话人分割性能[33]的好成绩。

（4）微软亚洲研究院提出了一种基于 UBM 的说话人实时分割算法[34,35]。该算法分为预分割和优化两步。在预分割阶段，根据每帧语音在 UBM 上的似然分的高低将该帧语音划分为可靠说话人语音帧、可疑说话人语音帧和非说话人语音帧；在优化阶段，使用递增说话人自适应（Incremental Speaker Adaptation，ISA）算法从可靠说话人语音帧上得到精确的说话人模型，并根据得到的模型对初始分割的结果做进一步判决。在 HUB4 英语新闻广播数据库上，误警率为 19.23%，漏检率为 13.65%。

（5）Delacourt 等 2000 提出了 DISTBIC 说话人分割算法[23]。该算法分为预分割和优化两步。在预分割阶段，采用 GLR 和 KL 距离作为距离度量准则进行初始分割；在优化阶段，使用 BIC 判断预分割结果中的相邻两个语音段是否属于同一个说话人，如果是则合并，否则保持不变。该算法在新闻语料和电话语料上都取得了不错的分割结果。

（6）北京大学信息科学技术学院智能科学系的视觉与听觉信息处理国家重点

实验室提出了一种基于集外说话人模型集上似然分向量的说话人分割算法[36]。该算法包括预分割、集外说话人模型打分和基于模型分数向量的聚类三部分，先将语音分割成每段只含有一个说话人的小段，然后进行集外模型打分并合并模型分数向量距离较小的段，采用重分割进一步提升性能，在 NIST 2003 双说话人识别数据库上取得了较好的分割效果。

（7）中国科学院自动化研究所高技术创新中心提出了一种基于熵的音频跳变点检测方法[37]用于广播电视环境下的说话人跟踪检测，切分后的语音片断通过说话人聚类来重新定位语音中说话人的变化点。新的语音片断，经过维纳滤波和 Pitch 端点检测，用于最终的说话人检测系统。在广播电视音频流上的错误率相对于 BIC 方法降低了 7.9%；目标说话人跟踪的综合统计的等错误率相对于基线系统降低了 9.5%。

（8）中国科学院声学研究所与中科信利实验室构建了一个完整的广播新闻语料识别系统（ThinkIT-BNR）[38]。该系统包括：音频匹配、音频自动分段、音频分类、说话人聚类、识别后处理和多阶段识别策略等多个模块，在新闻联播节目测试语料库上误识率为 10.14%，其中干净的播音员语音的误识率为 4.4%。

（9）浙江大学计算机科学与技术学院提出了一种多层次的说话人分割框架结构[39]，利用分层的结构特点，在各层引入辅助信息，利用语音和非语音的特征分布以及其突变规律，提出了基于决策树的语音分类检测方法，并应用 Anchor 模型[40]和基音信息进行说话人分割，在 YOHO 语音库[41]和 SRMC 语音库[39]都取得了不错的效果。

1.2.1.2　说话人分割聚类的常用特征

说话人分割聚类中使用的特征主要是低层声学特征，如线性预测倒谱系数（Linear Predictive Cepstrum Coefficient, LPCC）[42]、Mel 频率倒谱系数（Mel Frequency Cepstrum Coefficient，MFCC）[43-47]、感知线性预测系数（Perceptual Linear Predictive, PLP）[48,49]和线谱对（Line Spectral Pair，LSP）[22,35,50]等。也有一些研究人员利用时域参数和高层信息进行说话人分割和聚类，例如基音周期（Pitch）[22,50,51]、短时能量（Short-Time Energy, STE)[19]、过零率（Zero-crossing Rate，ZCR）[50]、短时能量比（Short-Time Energy Ratio，STER）[52]、频谱流（Spectrum Flux，SF）[52]、响度（Loudness 或 Energy）、共振峰（Formants）、谐波噪音比（Harmonics to Noise Ratio，HNR）和长时平均频谱（Long Term Average Spectrum, LTAS）等特征[53-56]。

此外，在设备条件允许下，利用麦克风阵列可以更好的对含有多个说话人的

语音进行处理[57]，但在现实中得到的绝大多数语音都是使用单一设备采集得到的而不是使用麦克风阵列，即使使用麦克风阵列录制的语音也很难获得麦克风阵列的详细信息，而且麦克风阵列的成本很高，因此在本文的研究中未考虑使用麦克风阵列的情况而只关注了基于距离度量或模型搜索的算法。

1.2.1.3　现有算法分析

基于距离度量的算法[13,17,21]的难点是阈值的确定，语音信号千差万别，不同说话人的语音段之间的距离可能很大也可能很小；属于同一个说话人的两段语音，由于说话人情感[58]变化、发音时所处的环境变化等的影响，距离也可能会较大，因此较难设定一个普适的阈值。

基于模型搜索的算法[7,10,28,29]的难点在于初始模型的选取，因为基于模型搜索的算法一般都需要从语音中选择语音段来估计可能的目标说话人模型，而如果用来训练初始模型的语音段的选择不恰当（即包含有多个人的语音），就会使得用于搜索的模型不正确或不精确，导致分割聚类的结果不好，而且该类算法复杂度高，需多次迭代计算，时间花销巨大。

两类算法还面临的一个共同的难点就是较短的两段语音之间距离的度量精度。在基于距离度量的算法中，度量相邻的两段语音之间的距离，根据距离大小判断两段语音是否属于同一说话人，每段语音不能太长，太长就可能包含一个以上的说话人[13,17,21]；在基于模型搜索的算法当中，选择一段语音来训练初始模型，同样也不能太长，搜索发音时刻是选取一小段语音并判断该段语音与初始模型的匹配程度[7,10,28,29]，归根到底也是判断两段语音之间距离，即训练初始模型的语音段和搜索语音段之间的距离。由于语音的数据较少而且没有任何可利用的先验知识，距离度量精度是较难得到保证的。显然，距离度量的精度直接影响分割聚类的性能。

1.2.2　说话人快速辨认的研究现状

1.2.2.1　常用的说话人识别方法

从说话人识别方法上说，常用的说话人识别方法可分为模板匹配法[59-62]、统计概率模型法、人工神经网络（Artificial Neural Network，ANN）法[63,64]和支持向量机（Support Vector Machine，SVM）[65-68]等，这些方法既能够用于说话人辨认也能够用于说话人确认。在文本无关的说话人识别领域，高斯混合模型和通用背景模型（Gaussian Mixture Model-Universal Background Model，GMM-UBM）[69]、高斯混合模型和支持向量机（Gaussian Mixture Model-Support Vector Model,

GMM-SVM）[70-72]、联合因子分析（Joint Factor Analysis, JFA）[73-77]是最常用的说话人识别方法，JFA 是在 GMM-UBM 基础上的改进算法。这些说话人辨认方法在一定条件下已经能达到很高的辨认准确率[69,70,73]，但是随着目标说话人数量的增多（几千、几万甚至更多）[4]以及输入数据的增多，这些方法的时间性能往往大幅下降难以满足应用需求，尤其是对于那些实时性要求较高的应用，例如大规模目标说话人检测系统。

1.2.2.2　常用的说话人辨认快速算法

基于 GMM-SVM 的说话人辨认系统[70]，一般首先利用 GMM-UBM 系统将待辨认语音在 UBM 上进行自适应得到高斯混合模型作为支持向量机的输入，自适应方法一般有最大后验概率（Maximum *A Posteriori*, MAP）[78-81]算法、最大似然线性回归（Maximum Likelihood Linear Regression, MLLR）[82-84]算法和本征音建模（EigenVoice Modeling）[73]算法等。MAP 需要估计出某一特定环境下的先验模型参数，从而对说话人模型进行相应的补偿；MLLR 假定用一小部分语音数据即可估计出训练环境与测试环境之间在模型参数上的差异，在此基础上，对说话人模型进行修正；EigenVoice 需要计算本征音因子的期望来估计本征音因子。这些算法的运算量大特别耗时，Liu 等 2002[85]使用分层结构的高斯混合模型（Hierarchical Gaussian Mixture Model, HGMM），Wang 等 2010[86]使用回归类树（Regression Class Tree, RCT）方法分别对 MAP 算法进行了加速，与基于 GMM-UBM 的说话人辨认系统相比，其运算速度仍有较大差距且较难进一步改进。因此，目前更多的说话人辨认快速算法是对基于 GMM-UBM 的说话人辨认系统进行改进。

在基于 GMM-UBM 的辨认系统中，运算速度主要受两个因素影响[86,87]。一是输入语音的特征向量的数量。每一帧特征向量都需要在 UBM 中挑选核心分布，一般是先在 UBM 所有的单高斯分布上计算似然分然后挑选似然分最高的前 N（N 为 4 或 5）个高斯分布作为核心分布[69]。而 UBM 作为通用背景模型是需要覆盖整个声学空间的，因此其混合数往往都比较大（1024 或 2048）[69]，而且似然分的运算主要为指数和对数运算，比较耗时，因此这部分运算量较大。二是目标说话人的数量。每一帧特征向量在 UBM 上挑选完核心分布之后，该帧特征需要在所有目标说话人模型中与 UBM 核心分布相对应的单高斯分布上计算似然分，这部分的运算量与目标说话人的数量成正比，当目标说话人规模很大时运算量也会很大。

目前已有的快速辨认算法可大体分为三种，一是快速挑选核心分布算法；二是下采样（Down-Sampling 或 Sub-Sampling）方法，压缩输入语音的特征向量数量，使用部分特征来代替全部特征；三是基于目标说话人聚类的剪枝算法，剪掉那些

与待辨认语音相似程度较低的目标说话人而只对保留的目标说话人进行说话人辨认，通过降低辨认的目标说话人数量减少运算时间。

（1）快速挑选核心分布算法。Auckenthaler 和 Mason 2011 提出了基于哈希表的核心分布快速挑选算法[88]，为 UBM 中的单高斯分布建立哈希表索引，利用哈希表查找的方式来快速挑选核心分布，核心分布的挑选速度提高了 10 倍而系统识别性能仅略微下降；Xiang 和 Berger 2003 提出了结构化高斯混合模型方法（Structural Gaussian Mixture Model，SGMM）[89]，识别时仅考虑树结构中各层分数最高的分布，把各层的似然分输入一个神经网络进行处理后得到一个单一分数用于说话人确认，在等错误率相对上升 5%的情况下识别速度提高 17 倍；熊振宇等 2006 提出了基于树形 UBM 的核心分布快速挑选算法（Tree Based Kernel Selection，TBKS）[90]，把 UBM 中的单高斯分布经过分层聚类后组织成树形结构，自顶向下搜索树形结构，每层剪枝掉似然分较小的节点仅保留似然分较高的节点，下层节点计算时仅对上层保留节点的子节点计算似然分，降低了单高斯分布的计算数量，从而提高挑选核心分布的速度，TBKS 算法在核心分布挑选速度提高了 14.8 倍情况下说话人辨认性能仅下降了不到 1%；Saeidi 等 2010 提出了基于排序高斯混合模型（Sorted Gaussian Mixture Model, Sort-GMM）的算法[91]，训练时使用排序函数来对 UBM 中的单高斯分布计算索引值并排序，识别时先使用排序函数求解特征的索引值，在排序的 UBM 中寻找索引值与该特征索引值相近的分布，计算似然分然后挑选核心分布，与结构化高斯混合模型方法[89]相比，该算法取得了更好的加速效果且同时性能下降很小。

（2）下采样算法。Mclaughlin 等 1999 提出了按间隔抽取特征向量进行识别的方法[92]；Pellom 和 Hansen 1998 提出了束搜索技术（Beam Search Technique，BST）[93]；熊振宇等 2005 提出了基于特征矢量重排序的剪枝算法（Observation Reordering Based Pruning, ORBP）[94]；Kinnunen 等 2006 提出了预量化（Pre-Quantizing, PQ）[95]算法等下采样算法。这类算法都利用了相邻语音帧的相关性较大，似然分之间的差异较小，且每帧语音计算的先后顺序与最终的似然分无关[92]。对语音进行下采样，如取采样间隔为 4 帧（即每 4 帧语音采样 1 帧）[94]得到 4 组语音，使用一组采样语音对目标说话人模型计算似然分来作为语音的似然分。

（3）基于目标说话人聚类的剪枝算法。Sun 等 2005 提出了分层结构的说话人辨认（Hierarchical Speaker Identification, HSI)[96]算法，Apsingekar 和 Leon 2007 提出了基于说话人模型聚类算法（Speaker Models Clustering, SMC）[4]。这两种算法在训练阶段利用某种聚类算法（如 ISODATA[97]或 K-Means[98,99]）对目标说话人的模型进行聚类，得到一组聚类中心，辨认时输入语音首先对聚类中心计算似然分，

然后选择得分最高的聚类中心所代表的那一类，对属于该类的目标说话人模型计算似然分进行辨认。Kinnunen 等 2006 提出了基于 PQ 的剪枝算法[95]，首先利用 PQ 进行下采样得到多组下采样语音，然后利用一组下采样语音对全部目标说话人模型计算似然分，剪枝似然分小于阈值或排名靠后的那些目标说话人；接下来增加一组采样语音对上一步剪枝后剩余的目标说话人模型计算并更新似然分然后剪枝，重复本步直到每组采样语音均被计算。

此外，还可以通过降低 UBM 的混合数或降低特征维数来提高基于 GMM-UBM 的说话人辨认系统速度的方法，但这类方法往往会因为模型或特征的精度不足引起辨认性能的较大幅度的下降，而且运算速度的提高幅度有限。

1.2.2.3　现有算法分析

目前的基于 GMM-UBM 的说话人快速辨认算法中，快速挑选核心分布算法性能很好，在提高挑选核心分布的速度的同时说话人辨认的性能下降很小甚至可忽略，但是当目标说话人的数量较多时，挑选核心分布的运算量仅占说话人辨认全部运算量的很小的一部分，该算法对整个辨认过程的加速作用较小；下采样方法在一定条件下能够提高辨认速度，但由于采样语音的似然分与全部语音的似然分之间存在一定的差异，算法的稳定性存在欠缺，而且因为用于辨认的特征往往是去除了静音的有效语音的特征，并不能保证相邻帧之间的相关性较大，因此算法的前提条件并不一定能够得到很好的满足；基于目标说话人聚类的剪枝算法对整个辨认过程的加速性能最好，缺点是目标说话人数量增大到一定程度后算法的辨认性能下降较明显，大规模目标说话人会影响聚类效果，使得各类之间的区分性降低，在两类或多类之间容易发生混淆的目标说话人数量较多，当待辨认语音中包含的目标说话人位于这些位置时，很可能会由于类的选择而无法被正确检测到，大规模目标说话人造成该类算法性能下降的根本原因在于剪枝后保留的目标说话人与聚类中心相似程度较高，而不能保证与待辨认语音相似程度较高，理想情况是挑选与待辨认语音相似程度较高的目标说话人。

1.2.3　大规模目标说话人检测的难点

大规模目标说话人检测需要将多说话人语音处理成多段单说话人语音然后再进行单说话人识别，一般来讲有以下难点：

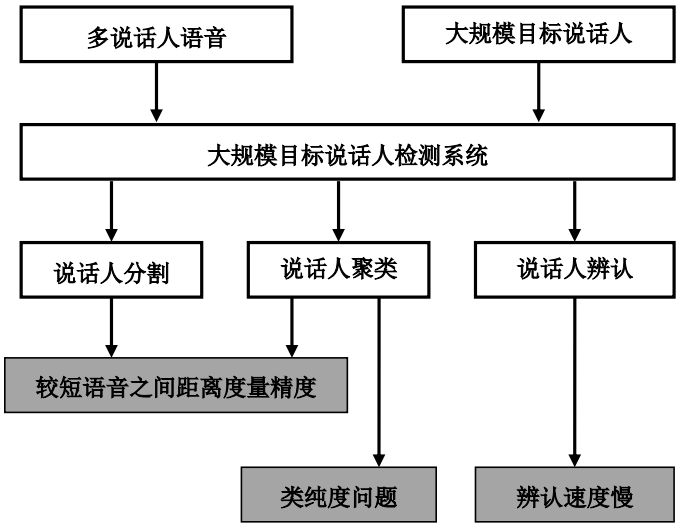


图 1.2 大规模目标说话人检测的难点问题

（1）说话人分割。说话人分割就是找到多说话人语音中的说话人转换（一个说话人停止说话而另一个不同的说话人开始说话）点，将多说话人语音分割成多段单说话人语音，说话人转换点的特征是该点两侧的语音分别属于不同说话人，说话人分割就是根据二者之间的差异来判断该点是否为说话人转换点。而实际的语音信号千差万别，两段属于不同说话人的语音之间的差异可能很大，也可能很小；两段属于同一个说话人的语音，由于发音长度的不同、发音内容的不同、说话人情感的变化以及说话时所处的环境变化等原因，使得二者之间的差异也很可能会有较大波动。度量两段语音之间差异没有任何的先验知识可以利用，而且每段语音不可能太长，太长的话很可能包含一个以上的说话人。根据以上分析看出，说话人分割是一个具有相当难度的问题。此外，说话人分割的结果直接影响到系统最终的说话人检测性能，如果有较多的漏检会直接影响后续的聚类以及辨认的性能；而如果有较多的误警会使分割得到的单说话人语音段长度过短也不利于聚类和辨认。漏检是指语音在某一时间点上发生了说话人转换却没有被分割，误警是指在某一时间点上语音没有发生说话人转换却被分割开。

（2）说话人聚类。在本文研究中，说话人聚类就是将说话人分割结果中属于同一个说话人的语音段聚成一类。度量两段语音之间的差异，根据差异大小来决定这两段语音是否属于同一类。而说话人分割之后得到的语音段的长度一般都比较短，在数据较少且没有先验知识的情况下度量两段语音之间的差异是具有相当难度的。而且本文研究的说话人聚类是无监督的聚类，不知道输入语音中说话人的数目，即聚类数，如果聚类结果中的类数少于说话人的数目，就很有可能使目

标说话人的语音被其他说话人的语音淹没；如果类纯度不好也会导致目标说话人的语音被其他说话人语音淹没或者混入其他说话人的语音，这都会对目标说话人检测的性能造成一定的影响，这些都说明说话人聚类是很有难度的，而且聚类效果会直接影响说话人辨认的性能。

（3）辨认速度。大规模目标说话人对说话人辨认速度的影响很大，辨认时间与目标说话人的数量成正比。说话人检测系统在现实应用中其目标说话人的数量往往特别巨大，同时说话人检测一般都有严格的速度要求。如何快速度量待辨认语音与目标说话人之间的相似程度，利用目标说话人剪枝来提高辨认速度，也是本文研究的一个难点问题。

综上所述，说话人分割、说话人聚类和快速辨认是目前大规模目标说话人检测技术无法回避的核心问题、关键问题、难点问题。

1.3　研究工作概述

1.3.1　研究思路

面向实际应用的大规模目标说话人检测系统，由于输入语音中大都包含多个说话人的语音，单说话人识别技术无法对其进行识别，先采用说话人分割聚类技术将多说话人语音转变为多段单说话人语音；在应用中，由于大规模目标说话人的影响以及说话人检测技术应用的速度要求，需要进行说话人快速辨认。

综上所述，本文采用的研究思路（如图 1.3 所示）为：首先对多说话人语音进行说话人分割处理，得到多段单说话人语音；分割之后得到的多段单说话人语音的每段长度可能会比较短，而较短的语音会使得说话人辨认由于数据较少而导致辨认性能较差或不稳定，接下来对分割后得到的语音段按照说话人的身份进行聚类处理，将属于同一说话人的语音合并以增加单说话人语音段的长度；使用说话人快速辨认算法提高说话人辨认速度。为了提高分割和聚类的性能，研究两段较短语音之间相似性的精确度量方法；为了降低聚类结果对辨认的影响程度，研究类纯度约束下的聚类方法；为了能够快速准确的进行说话人辨认对大规模目标说话人进行索引，并重构成易于进行目标说话人剪枝的结构，在此结构上研究快速辨认算法以提高辨认速度。

从以上分析可以看出，大规模目标说话人检测技术被分解为三个研究问题，即说话人分割、说话人聚类和说话人的快速辨认，对这三个问题的研究思路如下：

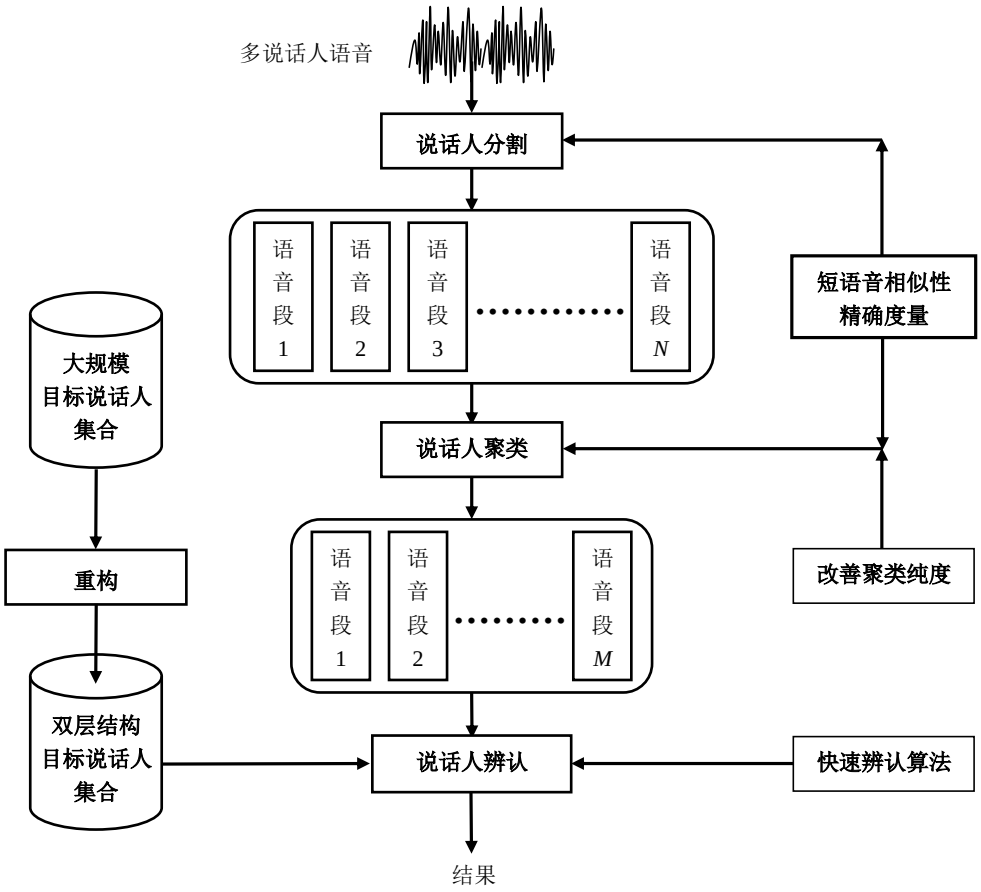


图 1.3 大规模目标说话人检测研究思路示意图

（1）说话人分割，基于模型搜索的分割算法的性能在很大程度上依赖于初始模型，如果训练初始模型的语音段选择不恰当（即包含有多个说话人的语音），就会使得用于搜索的说话人模型不正确或不精确，从而导致分割的性能不好，而且该类算法需要反复分割并更新模型使得算法速度很慢，同时考虑到说话人检测系统严格的速度要求，本文选择了基于距离度量的算法开展研究。基于距离度量的说话人分割聚类算法是根据左右相邻的两个分析窗内语音的距离来判断两个窗内的语音是否属于同一个说话人，其隐含条件是两窗内的语音最多发生一次说话人转换。由于输入语音中可能存在较短的语音段，这些短语音段限制了窗宽不能太长，过长的窗宽就很有可能在两窗语音内发生一次以上的说话人转换。而在较短的窗宽情况下，两窗语音之间的距离度量会由于数据量较少而容易导致不稳定与产生偏差。本文对说话人分割研究的出发点是尽可能提高在较短窗宽下的距离度量的准确性和稳定性，从而提高分割性能。常用的距离准则 BIC、GLR 和 KL 都先将每窗内的语音训练成模型，然后度量这两个模型之间的距离作为两窗语音

之间的距离，而模型的训练一般在数据量较大的情况才能保证训练得到的模型足够精确，因此会由于一窗语音的数据较少而导致模型训练不足，进而影响到距离度量的准确性和说话人分割的性能。本文改善分割性能的思路是增加可以利用的先验信息来提高距离度量精度，提出了两种距离度量准则。一是基于参考说话人模型的距离度量准则。一般来说，如果两窗语音属于同一个说话人，那么二者与同一个说话人模型的距离的差异较小，反之，如果两段语音与多个说话人模型的距离的差异都较小则说明二者属于同一个说话人的概率很高，如果两段语音与多个说话人模型的距离的差异较大则说明二者属于同一个说话人的概率较低。基于这一想法，设计了基于参考说话人模型的分割算法，训练多个参考说话人模型用来对声学空间进行分类描述，每个参考说话人模型代表一种典型说话人的声学特性，一窗语音与多参考说话人模型的距离则描述了这窗语音的声学特性。分别度量两窗语音与多参考说话人模型的距离得到两个距离矢量，然后将两个距离矢量之间的相似性作为分割用的距离度量准则。二是利用高层信息的基于音素识别和文本相关的距离度量准则，考虑到文本相关的说话人识别在语音很短的条件下（甚至仅有一两个字或词）都可以达到很高的识别性能，其根本原因在于识别语音和训练语音在内容上的存在着高度的一致性或相关性，本文中尝试借助音素识别技术获得语音中的音素信息，对输入语音进行音素识别得到音素序列，对两窗内相同的音素进行文本相关的说话人识别并将其识别结果作为距离度量，通过提升距离度量时内容的一致性和相关性来提升距离度量的准确性，从而进一步改善说话人分割性能。

（2）说话人聚类，在说话人数目未知的无监督聚类中，如果最终聚类的类数少于说话人的数目，就很有可能使目标说话人的语音被淹没在其他非目标说话人语音中，而如果聚类数多于说话人数目，发生目标说话人语音被淹没的概率会比较低。本文研究基于类纯度约束的说话人聚类算法，其基本思想是使聚类后达到最短辨认长度要求且类纯度较高的有效类尽可能多，降低不同说话人的语音被聚到同一类内的可能性。在说话人检测中，说话人聚类的目的是降低分割后得到的单说话人语音段的长度较短对说话人辨认性能的影响，因此只要单段语音的长度能够满足说话人辨认系统的最短长度要求即可，可允许同一说话人的语音聚成两类或多类。首先从待聚类的语音段中挑选最不可能属于同一说话人的两段语音作为初始的两类，挑选与这两类距离最近的一段语音，根据合并对类内离散度的影响决定合并还是增加新类，如果某一段语音达到了最短辨认语音长度要求，将该语音段单独归为一类不再参加后面的聚类过程，重复直至所有的语音都被处理。由于聚类不会使得在分割时产生的漏检错误得到消除，在聚类后进行了重分割处

理的话又会增加过多的处理时间，分割时尽量降低漏检率。

（3）快速辨认，在已有的说话人快速辨认算法中，快速挑选核心分布算法在大规模目标说话人条件下对说话人辨认整体的加速作用有限；下采样方法的采样语音的似然分与全部语音的似然分之间存在一定的差异，算法的稳定性存在欠缺，而且去除静音之后，相邻帧特征之间的相关性并不能保证较大；在大规模目标说话人条件下，基于目标说话人聚类的剪枝算法对辨认的加速作用最明显，现有的目标说话人剪枝算法 HSI[96]和 SMC[4]在目标说话人数量增大到一定程度后算法的辨认准确率会有明显下降，其根本原因在于其剪枝后保留的目标说话人与聚类中心相似程度高，但不能保证与待辨认语音相似程度高，本文提出了一种基于参考说话人和双层结构的快速辨认算法，借助参考说话人信息度量目标说话人和待辨认语音之间的相似程度，上层用来挑选参考说话人和快速粗剪枝，下层用来评估目标说话人和待辨认语音之间的相似程度，挑选与辨认语音最相近的一部分目标说话人进行辨认。

在对上述三个问题的研究基础上，可以构建一个完整的大规模目标说话人检测系统，虽然还存在许多需进一步研究的问题，但仍然可以在一定条件下进行实际应用，其中的技术也可独立应用，如快速辨认算法可单独应用于说话人辨认、基于参考说话人模型的距离度量也可应用于说话人确认等。

1.3.2　论文工作内容

论文的研究工作受多说话人语音和大规模目标说话人两个因素影响，面临说话人分割、说话人聚类和辨认速度三个方面的问题，如图 1.4 所示。

具体地说，作者的工作内容包括以下几个方面：

（1）针对说话人分割问题，为改善较短窗宽条件下两窗语音之间距离度量的精度和稳定性，提出了一种基于参考说话人模型（Reference Speaker Model，RSM）的说话人分割聚类算法。基于 RSM 的距离度量准则利用 RSM 能够较好的覆盖声学空间并能够描述各类典型说话人发音特性的特点，根据相邻两段语音在 RSM 上的似然分之间的差异来找出多说话人语音中的说话人转换点，并利用性别相关的 UBM 来提取语音的性别信息帮助确定说话人转换点，使用距离序列上的波峰波谷信息来进一步降低漏检和误警。研究了 RSM 的训练方法，以及基于 RSM 的两窗语音之间的距离度量方法，验证了参考说话人模型用来覆盖声学空间及描述声学特性的能力，在 NIST SRE 2002 说话人分割聚类评测数据库中的新闻采访语音库和电话对话语音库[5]上，与 BIC 和 GLR 度量准则对距离度量精度、分割性能进行了比较分析。

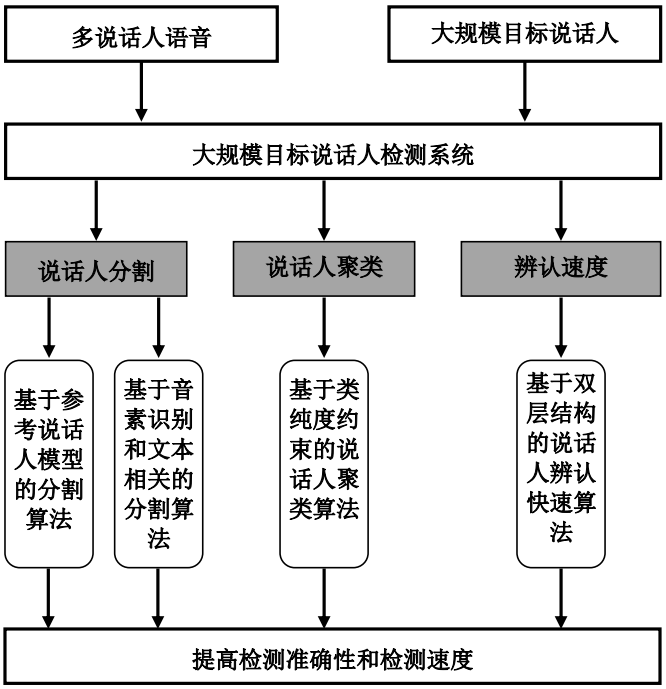


图 1.4 论文工作内容

（2）针对输入语音中说话人的平均发音时间较短（如电话交谈语音）的情况，提出了一种基于音素识别和文本相关的说话人识别的距离度量方法。该算法借助音素识别技术获取语音的音素信息，对输入语音进行音素识别得到音素序列，对音素序列分窗并对两窗内相同的音素进行文本相关的说话人识别并将识别结果作为距离度量。研究了如果两窗内不存在相同的音素的处理方法，分析了音素识别性能对于分割性能的影响，利用不同音素的说话人区分能力的差异对不同音素的距离的权重进行了调整，对不同音素距离的值域范围进行了归一化处理。在 TIMIT 数据库上，在说话人平均发音时间较短情况下，与 BIC、GLR 和 RSM 度量准则对距离度量精度、分割性能进行了比较分析。

（3）针对说话人聚类问题，提出了一种基于类纯度约束的说话人聚类方法。以聚类结果中达到最短辨认长度要求且类纯度高的有效类的数量最大为目标。首先分析了聚类结果对说话人检测的影响，聚类数大于说话人数的聚类结果对说话人检测的影响相对较小，聚类算法中假设待聚类的语音中至少存在两个不同的说话人。算法具体步骤如下：首先从语音段中挑选最不可能属于同一说话人的两段语音，分别挑选与这两段语音距离最小的一部分 RSM 用来进行距离度量，然后挑选与这两段语音中距离最近的一段语音，并进行合并判断，如果合并对于类内离散度影响较小则合并，否则该段语音作为新的一类，如果达到了辨认语音长度要

求，将该语音段单独归为一类不再参加后面的聚类过程，重复直至所有的语音都被处理。通过说话人聚类后，分割后的单说话人语音段的长度得到增加。输入的多说话人语音就改变为多段单人语音，得到了可直接用于单说话人辨认的语音。并与依赖阈值的分层聚类方法进行了实验比较和分析。

（4）针对辨认速度的问题，对已有的加速算法进行了分析，找到了大规模目标说话人对于 SMC 和 HSI 算法性能产生影响的根本原因，提出了基于参考说话人和双层结构的说话人快速辨认算法。算法步骤如下：将目标说话人模型、参考说话人模型组织成双层结构，利用待辨认语音与上层节点之间距离来度量待辨认语音和参考说话人之间的相似程度并对目标说话人进行粗剪枝，利用下层参考说话人快速挑选与待辨认语音最相近的一部分目标说话人。研究了如何构建双层结构，以及如何利用目标说话人或待辨认语音与节点的距离来度量二者相似程度的方法，并利用相似程度来动态挑选与待辨认语音相似的目标说话人，并分析了算法的加速性能。在中文语言资源联盟（Chinese Corpus Consortium, CCC）[100]的 VPR-2C2005-6000 数据库上对算法的辨认速度和辨认准确率等与 SMC[4]算法进行了比较分析。

1.4　论文的组织结构

本文的内容共六章，具体安排如下：

第 1 章是绪论部分，首先对大规模目标说话人检测技术的组成与发展做简要介绍，接着综述了其研究现状并分析归纳了大规模目标说话人检测存在的难点问题，并在此基础上阐述了研究的思路和工作内容。

第 2 章先介绍常见的基于距离度量的说话人分割算法，然后介绍为提高距离度量的准确性和稳定性提出的基于多参考说话人模型的距离度量准则和利用波峰波谷信息降低漏检和误警的方法，然后通过实验对提出的算法与 BIC、GLR 和 DISTBIC 算法进行了分析比较。

第 3 章先分析能否借助更多的信息来改善距离度量精度，介绍了两种使用高层信息的分割方法，然后介绍提出的基于音素识别和文本相关说话人识别的说话人分割算法，最后结合实验分析了该算法的有效性。

第 4 章先介绍说话人聚类的常用算法，接下来分析说话人聚类对于说话人检测性能的影响，然后介绍提出的基于类纯度约束的说话人聚类方法并结合实验分析算法性能。

第 5 章先介绍基于目标说话人聚类的剪枝算法并进行分析，然后介绍提出的用于 GMM-UBM 架构的说话人辨认系统、基于参考说话人和双层结构的目标说话

人剪枝算法，介绍了双层结构的设计方法和加速算法，最后在大规模目标说话人测试数据库上对各种算法进行了比较分析。

第 6 章是总结和展望部分，给出了论文所做工作和成果的总结，并指出了研究中存在的不足之处和对相关领域研究的展望。

第2章　基于参考说话人模型的说话人分割算法

语音信息随着信息技术的发展也越来越容易被人们所获得，例如电视节目录音、新闻采访录音、电话对话录音、网络聊天录音以及公安监听录音等等。这些语音信息通常以音频文件的形式存在，数量巨大并随着时间的推移不断的累积增多，但其中大多数语音为多说话人语音。为了能够更方便快捷地从海量语音资源检索到所需的说话人的信息、更充分地利用这些海量语音资源，需要对多说话人语音进行说话人分割处理得到多段单说话人语音，否则后续的单说话人识别无法进行。因此，说话人分割处理是说话人检测系统必须解决的问题之一。

本章内容安排如下：2.1 简单介绍常用的基于距离度量的说话人分割算法；2.2 介绍说话人分割算法的评价指标；2.3 介绍本文提出的基于参考说话人的说话人分割算法；2.4 给出几种算法在 NIST SRE 2002 说话人分割数据库上的实验结果和比较分析；2.5 是本章的小结。

2.1　基于距离度量的说话人分割算法介绍

基于距离的说话人分割算法中，一般选择某种距离度量准则来评价相邻两窗内语音的相似程度，也即根据距离的大小评价这两窗语音属于同一个说话人的概率，一般来讲距离大表明两窗语音属于同一个说话人的概率小，距离小则正好相反。两窗以一定的时间间隔从前向后移动并计算距离，直到到达语音段的终点，整段语音被全部处理后会得到一个距离序列，然后根据距离序列上峰值点以及相邻点的变化情况进行分割点判断。下面介绍几种说话人分割算法中常用的距离度量准则。

2.1.1　BIC距离度量

Chen 等人在 1998 年采用 BIC[13]（又被称为 Akaike 或 Rissanen 准则[101]）作为说话人分割的一种可分性度量方法。BIC 是一种基于模型复杂度（也即是模型参数的数量）惩罚的最大似然准则。给定一段输入语音其特征序列标记为 $F=\{f_1, f_2,..., f_N\}$，N 表示特征序列的数量。使用 F 生成的模型记为 M_A，$L(F|M_A)$ 是特征序列 F 在模型 M_A 上的似然分。在模型 M_A 上的 BIC 定义为：

$$BIC(M_A)=\log L(F\mid M_A)-\lambda\frac{m_A}{2}\log N \qquad (2\text{-}1)$$

其中，第一项表示输入语音与模型 M_A 的匹配程度，第二项表示模型 M_A 的复杂度惩罚分，其中 λ 是一个可调的平衡参数，m_A 是表示模型 M_A 中参数的个数。

BIC 不需要使用任何关于说话人的先验知识就能够描述模型与数据的匹配程度，因此它常被用来作为说话人分割的一种可区分性度量方法。BIC 在说话人分割中的应用示例如下（使用多维单高斯分布来表示说话人模型）：

设 $F_L=\{f_1,f_2,...,f_N\}$ 和 $F_R=\{f_{N+1},f_{N+2},...,f_{2N}\}$ 分别是输入语音中相邻的左右两窗语音的特征序列，此处的 N 是每窗语音中的特征数量，也就是窗宽的长度，做 H_0、H_1 两个假设：

H_0：如果相邻两窗语音属于同一个说话人 A，那么可以使用一个多维单高斯分布来描述该说话人，即 $F_A=\{f_1,f_2,...,f_N,f_{N+1},f_{N+2},...,f_{2N}\}\sim N(\mu_A,\Sigma_A)$。

H_1：如果两窗语音属于两个不同的说话人 L 和 R，即输入语音在时刻 N 发生了说话人转换，就可以用两个多维单高斯分布来描述说话人 L 和 R，即 $F_L\sim N(\mu_L,\Sigma_L)$和 $F_R\sim N(\mu_R,\Sigma_R)$。

μ 和 Σ 分别表示多维单高斯分布的模型参数，均值向量和协方差矩阵。H_0 和 H_1 之间的 BIC 距离：

$$R(N)=\frac{N_A}{2}\log|\Sigma_A|-\frac{N_L}{2}\log|\Sigma_L|-\frac{N_R}{2}\log|\Sigma_R| \qquad (2\text{-}2)$$

其中，N_A，N_L 和 N_R 分别表示说话人 A，L 和 R 对应的特征序列中的特征个数，$N_A=2N$，$N_L=N_R=N$。一般来说，说话人转换发生在 $R(N)$出现极大值的时刻 N_S。

$$N_S=\arg\max_N R(N) \qquad (2\text{-}3)$$

BIC 也可表示为：

$$\Delta BIC(N)=-R(N)+\lambda P \qquad (2\text{-}4)$$

其中，λ 是可调的平衡参数，P 定义如下：

$$P=\frac{1}{2}\left(D+\frac{1}{2}D(D+1)\right)\times\log N_A \qquad (2\text{-}5)$$

其中，D 表示特征的维数。如果$\Delta BIC(N)$小于 0，表示两窗语音属于两个不同的说话人，在时刻 N 发生了说话人转换；否则，表示两窗语音属于同一个说话人。

以 BIC 作为度量准则的说话人分割算法应用比较广泛，并且在不同类型的语音数据库上都取得了较好的分割效果[12,15,102]。

2.1.2　GLR距离度量

Gish 等 1991 提出 GLR 用于说话人辨认[17]，基本思想如下：给定两段输入语音，其特征序列分别记为 F_1 和 F_2。分别使用特征序列 F_1 和 F_2 生成模型 M_1 和 M_2；使用两段语音的特征 $F_A=\{F_1,F_2\}$ 生成模型 M_A。如果两段语音与两个模型匹配的更好，则两段语音属于不同的说话人的概率大；如果与一个模型匹配的好则两段语音属于同一个说话人的概率大。Delacourt 和 Wellekens 2000 将其用于说话人分割[23]，使用了多维单高斯分布来描述说话人模型，示例如下：

设 $F_L=\{f_1,f_2,...,f_N\}$ 和 $F_R=\{f_{N+1},f_{N+2},...,f_{2N}\}$ 分别是输入语音中相邻的左右两窗语音的特征序列，此处的 N 也代表分析窗的宽度，做 H_0、H_1 两个假设：

H_0：如果相邻两窗语音属于同一个说话人 A，那么可以使用一个多维单高斯分布来描述该说话人，即 $F_A=\{f_1,f_2,...,f_N,f_{N+1},f_{N+2},...,f_{2N}\}\sim N(\mu_A,\Sigma_A)$。

H_1：如果两窗语音属于两个不同的说话人 L 和 R，并在时刻 N 发生说话人转换，就可以用两个多维单高斯分布来描述这两个说话人，也就是说 $F_L\sim N(\mu_L,\Sigma_L)$和 $F_R\sim N(\mu_R,\Sigma_R)$。

H_0 和 H_1 之间的 GLR 距离定义：

$$GLR(N)=\frac{L(F_A\mid N(\mu_A,\Sigma_A))}{L(F_L\mid N(\mu_L,\Sigma_L))\cdot L(F_R\mid N(\mu_R,\Sigma_R))} \qquad (2\text{-}6)$$

其中，$L(F_A\mid N(\mu_A,\Sigma_A))$表示特征 F_A 在高斯分布 $N(\mu_A,\Sigma_A)$上的似然得分，$L(F_L\mid N(\mu_L,\Sigma_L))$ 表示特征 F_L 在高斯分布 $N(\mu_L,\Sigma_L)$ 上的似然得分，$L(F_R\mid N(\mu_R,\Sigma_R))$表示特征 F_R 在高斯分布 $N(\mu_R,\Sigma_R)$上的似然得分。

在实际应用中通常使用式（2-6）的 log 值来表示两段语音间的 GLR 距离：

$$d_R(N)=-\log GLR(N) \qquad (2\text{-}7)$$

GLR 的值越大（$d_R(N)$越小）表示假设 H_0 成立的概率越大，即两段语音属于同一个说话人的概率越大；*GLR* 的值越小（$d_R(N)$越大）表示假设 H_1 成立的概率越大，即两段语音属于不同的说话人的概率越大。一般来说，说话人转换发生在 $d_R(N)$出现极大值的时刻 N_S。

$$N_S = \arg\max_N d_R(N) \tag{2-8}$$

GLR 准则不仅可用于说话人分割[103,104]，同时也可用于说话人辨认和说话人确认，都取得了不错的效果[17]。

2.1.3　KL距离度量

设 $F_L = \{f_1, f_2, ..., f_N\}$ 和 $F_R = \{f_{N+1}, f_{N+2}, ..., f_{2N}\}$ 分别是输入语音中相邻的左右两窗语音的特征序列，用两个多维单高斯分布来描述这两窗语音，即 $F_L \sim N(\mu_L, \Sigma_L)$ 和 $F_R \sim N(\mu_R, \Sigma_R)$。做 H_0、H_1 两个假设：

H_0：如果相邻的两窗语音属于同一个说话人 *A*，则 $N(\mu_L, \Sigma_L)$ 与 $N(\mu_R, \Sigma_R)$ 两个高斯分布之间的 KL 距离较小。

H_1：如果相邻的两窗语音属于两个不同的说话人 *L* 和 *R*，则 $N(\mu_L, \Sigma_L)$ 与 $N(\mu_R, \Sigma_R)$ 两个高斯分布之间的 KL 距离较大。

对于 $N(\mu_L, \Sigma_L)$和 $N(\mu_R, \Sigma_R)$两个多维单高斯分布，二者之间的 KL 距离[89]：

$$\begin{aligned} KL(N) = &\frac{1}{2}(\mu_R - \mu_L)^T \left(\Sigma_L^{-1} + \Sigma_Y^{-1}\right)(\mu_R - \mu_L) \\ &+ \frac{1}{2} tr\left(\left(\Sigma_L^{1/2}\Sigma_R^{-1/2}\right)\left(\Sigma_L^{1/2}\Sigma_R^{-1/2}\right)^T\right) \\ &+ \frac{1}{2} tr\left(\left(\Sigma_L^{-1/2}\Sigma_R^{1/2}\right)\left(\Sigma_L^{-1/2}\Sigma_R^{1/2}\right)^T\right) - D \end{aligned} \tag{2-9}$$

其中，$tr(\cdot)$表示矩阵的迹运算，*D* 为特征的维数。KL 距离也可以使用式（2-10）进行简化计算[2]：

$$\begin{aligned} KL(N) = &\frac{1}{2}(\mu_R - \mu_L)^T \left(\Sigma_L^{-1} + \Sigma_R^{-1}\right)(\mu_R - \mu_L) \\ &+ \frac{1}{2} tr\left(\left(\Sigma_L - \Sigma_R\right)\left(\Sigma_R^{-1} - \Sigma_L^{-1}\right)\right) \end{aligned} \tag{2-10}$$

式（2-10）中等号右边的第一项的值决定于均值和方差，第二项的值决定于方差，也有的研究人员简化了式（2-10）而只使用其中的第二项来计算 KL 距离[2]，即：

$$KL(N)=\frac{1}{2}tr\left((\Sigma_L-\Sigma_R)(\Sigma_R^{-1}-\Sigma_L^{-1})\right) \tag{2-11}$$

在基于距离度量的说话人分割算法中，相邻两窗语音的 KL 距离越大则这两窗语音属于不同的说话人的概率越大。一般来说，说话人转换发生在 $KL(N)$出现极大值的时刻 N_S。

$$N_S=\arg\max_N KL(N) \tag{2-12}$$

同 GLR 准则一样，KL 距离准则不仅可用于说话人分割，同时也可用于说话人辨认和说话人确认，都取得了不错的效果。

2.2　说话人分割算法的评测指标

说话人分割算法常用的评测指标主要有误警率（False Alarm Rate，FAR）和漏检率（Miss Detection Rate，MDR）。

误警是指给出的分割点实际上并不存在，即分割点左右相邻的两段语音是属于同一个说话人。

FAR 的定义为：

$$FAR=\frac{\text{误警的个数}}{\text{真实分割点的个数}+\text{误警的个数}}\times 100\% \tag{2-13}$$

漏检是指实际存在的说话人分割点没有被检测出，即分割后的语音段里含有多个说话人。

MDR 的定义为：

$$MDR=\frac{\text{漏检的个数}}{\text{真实分割点的个数}}\times 100\% \tag{2-14}$$

一般来说，FAR 越低，MDR 就相对越高，反之亦然。对一个说话人检测系统来说，MDR 的危害性要远远大于 FAR，但不是说 MDR 越低越好，这是因为 MDR 越低，FAR 相应就越高，分割后得到的语音段平均长度也就越短，不利于后面的说话人聚类和辨认处理。

召回率和精确率也是常用的说话人分割评价指标等。召回率（Recall Rate）指检测到的正确的说话人转换点在所有真实的切换点中所占的比例。精确率（Precision Rate）指检测到的正确的说话人转换点在检测到的所有说话人转换点中所占的比例。

在本文中使用误警率和漏检率来评价分割算法的性能。

2.3　基于参考说话人模型的说话人分割算法

2.3.1　问题的提出

在基于距离度量的说话人分割聚类算法中，根据左右相邻的两窗语音之间的某种距离来判断两窗语音是否为同一个说话人，在前文中分析过窗宽不宜太长，否则的话两窗内就很有可能发生多于一次的说话人转换导致算法错误；而在较短的窗宽条件下直接度量两窗语音的距离会由于数据少而容易产生较大的偏差或不稳定。常用的距离准则 BIC、GLR 和 KL 距离都是使用一窗语音生成模型（多维单高斯分布）来描述说话人的特性，然后使用这个模型进行距离度量，而由于一窗语音的数据较少势必会导致生成的模型欠缺准确性，此外单高斯分布对说话人的描述也不够精确（对说话人精确描述一般使用高斯混合分布），这就会导致后续的度量容易产生偏差以及不稳定，只有在某些条件下（如平均说话人转换时间较长同时取较长的窗宽）分割算法才能够保持一个较好的性能。为了改善说话人分割性能，必须首先提高距离度量在较短的窗宽条件下的准确性和稳定性，本章中作者提出了一种基于参考说话人模型的距离度量准则，并以此为基础进行说话人分割。

2.3.2　基本思想

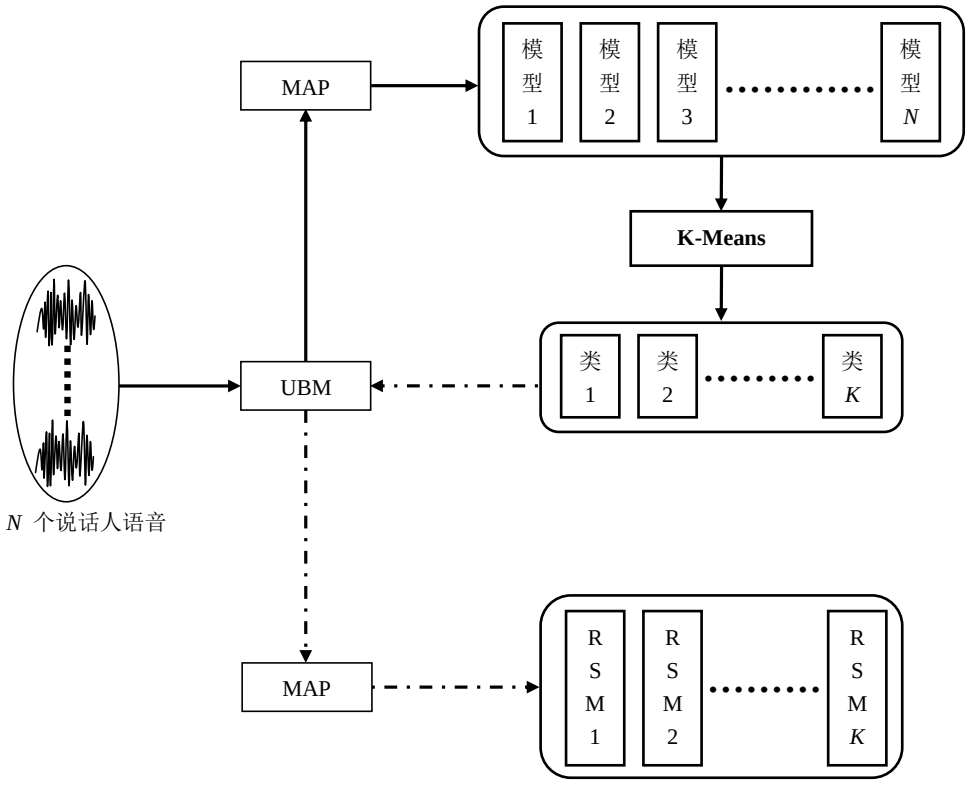
如果两窗语音属于同一个说话人，那么二者与同一个说话人模型的距离的差异较小，反之，如果两段语音与多个说话人模型的距离的差异都较小则说明二者属于同一个说话人的概率很高，如果两段语音与多个说话人模型的距离的差异较大则说明二者属于同一个说话人的概率较低。对这些说话人模型的要求是覆盖的声学空间足够广且足够精确，这样才能保证每窗语音都有对应的距离较近的说话人模型，本文将其称为参考说话人模型（Reference Speaker Model，RSM）。训练覆盖整个声学空间的多个参考说话人模型，每个 RSM 代表了一种典型的说话人发音特性，一窗语音与多个 RSM 的距离则能够用来描述其声学特性。分别度量两窗语音与多个 RSM 的距离得到两个距离向量，这两个距离向量的距离可以

描述两窗语音属于同一个说话人的概率大小。距离小则两窗语音属于同一个说话人的概率大，距离大则属于不同的两个说话人的概率大。

2.3.3　算法描述

首先选择一定数量的说话人模型，这些说话人模型用来近似模拟整个声学空间，然后通过矢量量化（Vector Quantization, VQ）方法来训练参考说话人模型。选择合适的窗宽和窗移，计算相邻两窗与多参考说话人模型的距离得到距离序列，计算两个距离向量之间的距离。两个分析窗同时向后移动并计算距离，直到整段语音被处理完毕后得到一个距离序列。根据距离序列上极值及变化判断分割点。

2.3.3.1　基于 VQ 的参考说话人模型训练方法



2.1 参考说话人模型训练示意图

（1）选择尽可能多的不同说话人来覆盖声学空间，这些说话人与训练 UBM 的语音以及说话人检测的目标说话人语音不存在重叠，且满足性别均衡、信道均衡和语言均衡等条件。给定一个说话人 s 的一段语音，用 MAP 算法[80]从 UBM 上自适应得到 s 的模型 $M(s)$，方差和权重都保持不变，而仅自适应均值。假设特征的维数为 D，给定一个特征向量 X，高斯混合模型的似然函数：

$$p(X \mid \mathrm{GMM}) = \sum_{i=1}^{M} w_i g_i(X) \tag{2-15}$$

其中，M 是 GMM 中高斯混合的个数，w_i 是第 i 个高斯混合的权重且满足：

$$\sum_{i=1}^{M} w_i = 1 \tag{2-16}$$

$g_i(\cdot)$ 是期望为 μ_i，协方差矩阵为 Σ_i 的高斯混合的概率密度函数：

$$g_i(X) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)\right\} \tag{2-17}$$

在本文中高斯混合模型中的协方差矩阵均使用了对角协方差矩阵。假设对说话人 s 的语音提取特征后得到 T 帧特征，$X = \{x_1, x_2, ..., x_T\}$，均值自适应公式：

$$\hat{\mu}_i = \alpha_i E_i(X) + (1-\alpha_i)\mu_i^{\mathrm{UBM}} \tag{2-18}$$

$$E_i(X) = \frac{1}{n_i}\sum_{t=1}^{T} p(i \mid x_t) x_t \tag{2-19}$$

$$p(i \mid x_t) = \frac{w_i \cdot g_i(x_t)}{\sum_{k=1}^{M} w_k \cdot g_k(x_t)} \tag{2-20}$$

$$n_i = \sum_{t=1}^{T} p(i \mid x_t) \tag{2-21}$$

$$\alpha_i = \frac{n_i}{n_i + \gamma} \tag{2-22}$$

其中 $\hat{\mu}_i$ 表示第 i 个混合自适应之后的均值，$g_i(x_t)$为第 i 个高斯混合的概率密度函数，x_t 表示第 t 帧语音特征。式（2-22）中的γ是模型先验分布的一个重要参数，它控制着自适应对先验信息 μ_i^{UBM} 的依赖程度。γ取值越大自适应后的说话人模型参数越接近于 UBM 的参数；γ取值越小自适应后的说话人模型参数则更多地由训练语音决定。在训练数据有限的情况下，一般采用较大的γ值，在本文的实验中γ设为 16。

（2）K-Means 聚类

模型间距离的度量。使用 KL 距离来度量说话人模型之间的距离，两个 GMM 之间的 KL 距离基于两个多维单高斯分布之间的 KL 距离进行计算，假设两个目标说话人的 GMM 分别为λ_1和λ_2，二者之间的 KL 距离计算公式如下式：

$$KL\left(\lambda_1,\lambda_2\right)=\sum_{i=1}^{M}w_i KL\left(g_i^1,g_i^2\right) \tag{2-23}$$

式(2-23)中，g_i^1，g_i^2分别为说话人模型λ_1和λ_2的第 i 个单高斯分布，$KL\left(g_i^1,g_i^2\right)$为使用式（2-10）计算的$g_i^1$与$g_i^2$之间的 KL 距离，$w_i$为 UBM 中第 i 个单高斯分布的权重。

使用最小最大方法从 N 个说话人模型当中选择 K 个作为初始的聚类中心。式（2-24）定义了一个 GMM 与一个 GMM 集合 S 的距离，其中 S 代表一个说话人模型集合即 GMM 集合，λ_i代表了一个 GMM。

$$d\left(\lambda_i,S\right)=\min_{j\in S}KL\left(\lambda_i,\lambda_j\right) \tag{2-24}$$

定义两个 GMM 集合 S_1 和 S_2，S_1 的初始状态为用来训练参考说话人模型的 N 个说话人的 GMM，S_2 的初始状态为空。首先利用 UBM 选择第一个聚类中心，计算 $d\left(\mathrm{UBM},S_1\right)$并将满足式（2-25）的 λ_j 作为第一个初始中心，将 λ_j 加入集合 S_2 并将其从集合 S_1 中移除；然后利用式（2-26）选择下一个初始中心，并更新集合 S_1 和 S_2，重复本步骤直到集合 S_2 中的初始中心达到 K 个。

$$j=\arg\min_{j\in S_1}KL\left(\mathrm{UBM},\lambda_j\right) \tag{2-25}$$

$$\lambda=\arg\max_{i\in S_1}d\left(\lambda_i,S_2\right)=\arg\max_{i\in S_1}\min_{j\in S_2}KL\left(\lambda_i,\lambda_j\right) \tag{2-26}$$

式（2-26）定义的是寻找 GMM 集合 S_1 中与 GMM 集合 S_2 距离最大的 GMM 的方法。

分别计算每个说话人模型与 K 个中心的 KL 距离，并将这 K 个距离排序，将这个说话人划分到与其 KL 距离最小的那个中心；迭代计算更新聚类中心直到 K-Means 聚类算法停止。

（3）K-Means 聚类后得到的 K 个不同的类，在属于同一类的说话人模型当中，挑选与类中心距离最近的前 50%的说话人的语音，将其合并然后在 UBM 上利用 MAP 算法仅自适应均值后得到的说话人模型即为参考说话人模型。

（4）由于性别信息对于说话人分割是十分直接有效的信息，在这 K 个参考说话人之外，增加性别信息参考说话人模型 F 和 M，F 代表女声参考说话人和 M 代表男声参考说话人。F 和 M 是两个性别相关的 UBM，分别由女声语音和男声语音使用 EM 算法迭代计算获得[105]。

2.3.3.2　距离度量

假设有一段语音 $X=\{x_1,x_2,...,x_T\}$，T 为特征的个数。首先使用参考说话人模型来描述语音的声学特性。本文中使用语音对参考说话人模型的似然分来描述语音该参考说话人模型的距离或相似程度。同 GMM-UBM 的似然分计算相同[69]，一帧语音首先在 UBM 上挑选核心分布并计算似然分，本文中核心分布的个数取 4。然后计算该帧语音在参考说话人模型中与 UBM 核心分布相对应的高斯分布上的似然分。

假设一帧语音在 UBM 上得分为 S_U，在某参考说话人模型上的得分为 S_{RS}，二者相减即为该帧特征在此参考说话人模型上的最后得分。语音 X 的在某参考说话人模型上的最终得分的计算公式如下：

$$S=\frac{1}{T}\sum_{i=1}^{T}\left(S_{RS}^{i}-S_{U}^{i}\right) \tag{2-27}$$

$$S_{U}^{i}=\log\left(p\left(X\mid \mathrm{UBM}\right)\right)=\sum_{j=1}^{N}w_{Top(j)}g_{Top(j)}^{U}\left(X\right) \tag{2-28}$$

$$S_{RS}^{i}=\log\left(p\left(X\mid \mathrm{RSM}\right)\right)=\sum_{j=1}^{N}w_{Top(j)}g_{Top(j)}^{RS}\left(X\right) \tag{2-29}$$

S_{RS}^{i} 和 S_{U}^{i} 分别是第 i 帧特征在参考说话人模型和 UBM 上的似然分，$1\le i\le T$，$Top\left(j\right)$ 指 UBM 中似然分最高的第 j 个单高斯分布的索引，N=4。

由于不同的参考说话人模型的似然分范围可能差异较大，而过大的差异对似然分向量的距离计算的准确性会产生影响，本文对似然分范围进行了规整，规整公式如下：

$$S_{norm}=\frac{1}{1+e^{-s}} \tag{2-30}$$

式(2-30)是利用 *Sigmoid* 函数将参考说话人模型的似然分范围规整为(0,1)。

性别参考说话人似然分的计算稍有不同，仅计算每一帧语音对性别参考说话人模型 F 或 M 中的核心分布上的似然分并将其作为性别信息。

计算语音 X 对所有参考说话人模型和性别信息上的似然分得到距离向量 $L_v(X)$。

$$L_v(X)=\left[S_{X_i}\right]^T, i=1,2,...,R,F,M$$

S_{X_i} 代表语音 X 对第 i 个参考说话人模型上的似然分，S_{X_F} 和 S_{X_M} 分别是对性别信息参考模型的似然分。将另一段语音 Y 的距离向量记为 $L_v(Y)$。

$$L_v(Y)=\left[S_{Y_i}\right]^T, i=1,2,...,R,F,M$$

计算距离向量 $L_v(X)$ 和 $L_v(Y)$ 之间的距离作为这两段语音之间的距离，本文中使用了相关系数作为距离度量。距离公式为：

$$d(X,Y)=1-C_{XY}/\delta_X*\delta_Y \tag{2-31}$$

其中，C_{XY} 是距离向量 $L_v(X)$ 与 $L_v(Y)$ 之间的协方差系数，δ_X 和 δ_Y 分别是似然分序列 $L_v(X)$、$L_v(Y)$ 的标准差。

$$C_{XY}=E\left(\left(X-E(X)\right)\left(Y-E(Y)\right)\right) \tag{2-32}$$

$$\delta_X^2=\frac{1}{K+1}\sum_{i=1}^{K+2}\left(S_{X_i}-S_{x_i}^m\right)^2,\ S_{X_i}^m=\frac{1}{K+2}\sum_{i=1}^{K+2}S_{X_i} \tag{2-33}$$

如果式（2-31）的值越小表示两段语音属于同一说话人的概率越大，值越大表示两段语音属于不同说话人的概率越大。

2.3.3.3　分窗计算

图 2.2 是分窗的示意过程。图 2.2 中 *d* 为窗移，即分割的精度。Window1 和 Window2 是相邻的两个分析窗，*t* 为相邻两窗的边界位置所对应的时刻。分窗有窗宽和窗移两个参数。一般来说，窗宽不能够取太长，否则两窗内发生一次以上的说话人转换可能性会很高，算法的前提条件得不到满足。窗移表示说话人分割的精度，窗移 *d* 越小，则精度会越高，而耗时也就越长；反之窗移越大，分割精度就低，分割耗时也就越短。可根据输入语音的类型不同，选择不同的窗长和窗移。对于说话人一次发音平均时间较长的新闻采访语音，窗长和窗移可以适当大些；而对于语音中含有较多短发音的电话交谈语音或对话语音，窗长和窗移一般略小。

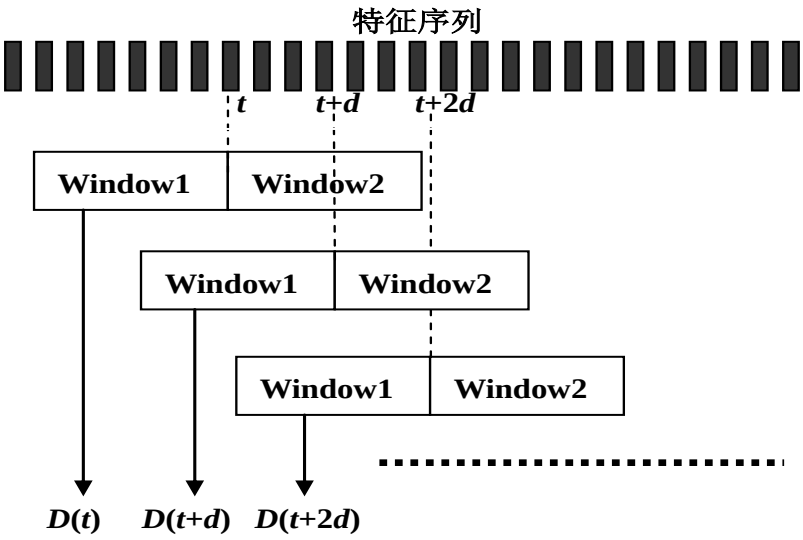


图 2.2 分窗示意图

图 2.2 中 *D*(*t*)表示以 *t* 时刻为边界的相邻两窗语音之间的距离，将两窗沿着时间轴从前向后同时移动并计算距离，最后得到一组距离值可形成距离序列。距离序列以图形表示即为距离变化曲线，参见图 2.3。

2.3.3.4　分割点判断

基于距离度量的说话人分割算法中的一个难点是判断分割点的阈值的确定，因为在实际应用中语音信号千差万别，两段属于不同说话人的语音之间距离有的可能很大，而有的可能很小；两段属于同一个说话人的语音，由于发音长度、发音内容、说话人情感差异以及说话时所处的环境等因素的影

响，使得计算出的距离值很可能会有较大的波动。因此，基于距离度量的说话人分割算法一般设定动态的阈值。假设某距离度量准则距离较大表示两段语音相似程度较低，距离较小表示两段语音相似程度高，那么再距离曲线上的每个极大值点是分割点的概率较大，有可能是说话人分割点，本文利用距离序列的方差来判断一个极大值点是否为一个说话人分割点[23]。

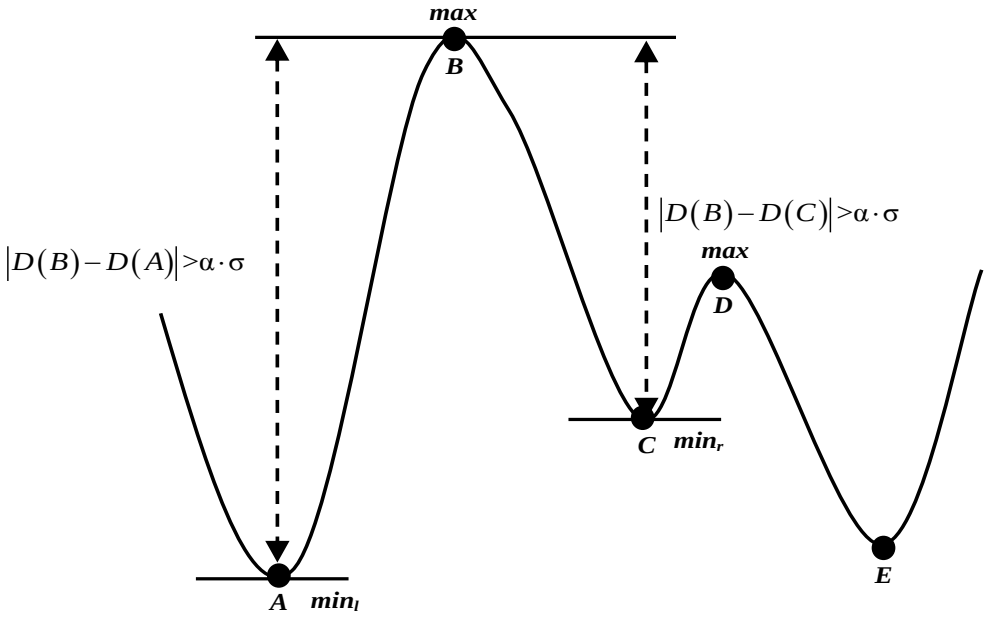


图 2.3 距离变化曲线

图 2.3 给出了一段距离变化曲线上的几个极值点，*A*、*B*、*C*、*D* 和 *E*。*B* 和 *D* 是两个局部极大值点，可能为分割点，由于 *B* 点与其相邻的局部极小值点的距离之差与 *D* 点相比较大，因此 *B* 点相对于 *D* 来说，成为分割点的可能性更大。*A* 和 *C* 分别是 *B* 左、右两侧最近的两个局部极小值点。

假定该距离序列服从高斯分布，高斯分布的方差表示了数据的离散程度，使用距离序列的方差来确定最后的分割点判决阈值。给定距离序列上的一个极大值点和它的左右相邻的两个极小值点，如果极大值点与其相邻的某一个极小值点对应的数值之差大于给定的阈值，那么该极大值点就是一个可能的分割点；否则，该极大值点不是分割点。

判决阈值一般取 $\alpha\cdot\sigma$，其中 α 是一个可调节的参数，σ 是在距离序列上求出的标准方差，公式如下：

$$t(max)=\begin{cases}\text{分割点，} & \text{若}\left|max-min_l\right|>\alpha\cdot\sigma \\ & \text{且}\left|max-min_r\right|>\alpha\cdot\sigma \\ \text{不是分割点，} & \text{否则}\end{cases} \qquad (2\text{-}34)$$

其中，max 是距离序列中的一个极大值点；min_l 和 min_r 分别是与 max 最近的左、右两个局部极小值点；$t(max)$表示 max 所对应的时间点。

2.3.3.5　利用波峰波谷进行漏检误警判断

通过步骤 2.3.3.4 进行分割点判断后，得到初步的分割结果，但此时的分割结果一般会存在一定的误警和漏检，误警分割点可以通过后续的聚类处理得到一定的消除，而漏检则是聚类处理无法消除的。为了减少漏检和误警，本文利用距离序列上的波峰波谷信息做进一步判断。

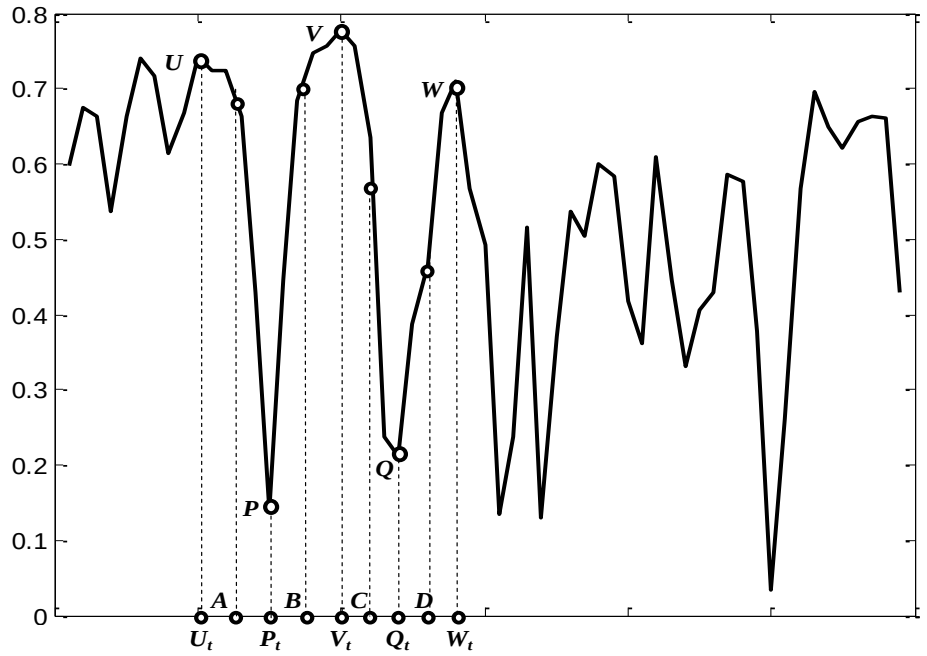


图 2.4 利用波峰波谷信息

图 2.4 中，横坐标表示时间，纵坐标表示两窗语音之间的距离。假设 U，V 和 W 是三个相邻的说话人分割点也是局部极大值点即波峰，而是 P 和 Q 分别是 UV 和 VW 中间的两个局部极小值点即波谷。U_t，P_t，V_t，Q_t 和 W_t 分别是 U、P、V、Q 和 W 在水平时间轴上的投影。

$$A = (U_t + P_t)/2,\ B = (P_t + V_t)/2$$

$$C = (V_t + Q_t)/2,\ D = (Q_t + W_t)/2$$

计算语音段 AB 与语音段 CD 之间的距离度量记做 D_p，计算语音段 U_tA 与语音段 BV_t 之间的距离度量记做 D_v。AB 与 CD 两个语音段是语音段 UV 中最稳定的部分，A 时刻和 B 时刻之间、C 时刻和 D 时刻之间存在说话人转换概率很小。U_tA 与 BV_t 是语音段 UV 中最不稳定的部分，U_t 时刻和 A 时刻之间、B 时刻和 V_t 时刻之间存在说话人转换的概率较大。

如果 $D_p < \beta$，则 V 可能不是一个真正的说话人分割点，相邻的两段语音属于同一个说话人，可以删除分割点 V。

如果 $D_v > \rho$，则在 U_t 与 V_t 之间可能存在一个说话人转换，采用小窗移进一步分析该段语音。

β 和 ρ 是两个预定义的阈值，在开发集上实验得到。在实验中 β = 0.3，ρ= 0.4 实验结果比较理想。

2.4　实验结果与分析

2.4.1　实验数据和设置

本实验中使用了 GMM-UBM 进行研究，UBM 是选择 NIST SRE 2004[5] 数据中的 368 位女说话人、248 位男说话人的语音数据使用 EM 算法[105]训练得到的，每人一段大约 5 分钟的语音，数据总量大约为 51 小时；训练参考说话人模型数据选择了 NIST SRE 2005、2006 和 2008 数据中的 1277 位说话人的语音[5]，每位说话人语音的长度大约为 60 秒，性别信息参考说话人模型分别使用其中的男声数据和女声数据训练两个性别相关的 UBM[105]；从 NIST SRE 2008 数据中挑选了男女各 100 人用于参数选择比较实验，每人 3 段语音，长度分别为 1 秒、2 秒和 5 秒，对应分割中的较短窗宽、一般窗宽和较长窗宽；训练 UBM 使用的说话人集合，训练参考说话人模型的说话人集合以及参数选择比较实验用的说话人集合均没有重叠；说话人分割实验选用的是 NIST SRE 2002 说话人分割数据库中的 BNEWS 数据库（新闻采访语音）和 Switchboard 数据库（电话对话语音，简写做 SWBD）[5]。BNEWS 数据库是新闻采访语音，共有 76 个语音文件；SWBD 数据库是双人电话交

谈语音，共有 199 个语音文件。所有的语音均处理成采样频率 8,000Hz，采样精度 8bit。

实验中使用 MFCC 作为特征[43]，前端处理使用的帧长为 20 毫秒，帧移为 10 毫秒，预加重系数为 0.97，窗函数为汉明窗（Hamming Window），每帧语音使用的 256 点的快速傅立叶变换，截止频率为 300Hz~3,400Hz，Mel 滤波器组的个数为 30，滤波器组等带宽，中心频率等间隔分布，滤波器组的输出经过 Log 压缩和离散余弦变换之后，得到相应的 16 维特征系数，并提取相应的一阶差分 16 维，共 32 维特征系数。使用基于能量的语音活动检测（Voice Activity Detection，VAD）算法去除静音，在训练 UBM 和参考说话人模型时使用去除静音后的特征；在说话人分割时先使用 VAD 进行静音标注，分割过程中的分窗使用未去除静音的特征计算时间信息，距离度量时使用去除静音的特征进行计算，最后利用 VAD 的静音信息辅助确认分割点。最后对去除静音之后的特征进行倒谱均值减和倒谱方差归一化（Cepstral Mean Subtraction-Cepstral Variance Normalization, CMS-CVN）[45]处理后得到最后的语音特征序列。本文后面章节里所用到的语音特征提取参数均按上述方法进行提取，以后不再赘述。

实验中评测了三种系统：基于 BIC 的说话人分割系统[13]；基于 GLR 的说话人分割系统[17]；DISTBIC 分割系统[23]和基于参考说话人模型（RSM）的说话人分割系统。

2.4.2　实验结果与分析

实验一　RSM 数量的选择实验

实验目的是为了选择合适的 RSM 数量，RSM 是使用 VQ 算法训练获得，理论上 RSM 数量越多越好，但 RSM 数量过多会由于训练数据稀疏而导致性能下降。本文中首先利用实验来选择 RSM 的数量。

实验数据是使用参数选择数据进行说话人确认实验，实验中每次使用相同长度的两段语音，利用参考说话人模型计算两段语音之间的距离，最后根据算法的说话人确认性能来选择合适的 RSM 数量。

评价说话人确认系统的性能有错误接受率（False Acception Rate，FAR；也被称为 False Alarm Rate）和错误拒绝率（False Rejection Rate，FRR；也被称为 Miss Probability）两项参数[5]。前者指的是系统对假冒者的接受性能，该值越低说明系统越安全，不易被闯入；后者指的是系统对真实说话人的拒绝性能，该值越低说明真实说话者越容易进入系统。一般来说判决阈值越低，

系统的 FRR 越低则相应的 FAR 就越高；阈值越高，系统的 FRR 越高则相应的 FAR 就越低。也就是说，FAR 和 FRR 都是判决阈值的函数，这两个函数在值域相交的点称为等错误率（Equal Error Rate，EER）点。通常人们希望系统的等错误率尽可能低，也就是 FAR 和 FRR 相等时的值尽可能小。

检测错误权衡曲线（Detection Error Trade-offs Curve，DET Curve）是另一种常用的评价方法[106]，检测错误权衡曲线反映 FAR 和 FRR 之间的关系。DET 曲线越接近原点，系统的识别性能越好。

NIST[5]还定义了 FAR 和 FRR 的加权和函数，检测代价函数（Detection Cost Function，DCF）。针对不同的应用背景对 FAR 和 FRR 定义不同的权重（代价），并用最小 DCF（Minimum DCF，*mDCF*）来表示系统能够取得的最优性能。对于实际应用的系统，*mDCF* 要比 EER 更有意义。在给定不同错误率权重（代价）下，*mDCF* 越小，系统的实际应用性能越好。DCF 的定义为：

$$C_{Det} = C_{Miss} \times P_{Miss} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm} \times \left(1 - P_{Target}\right) \qquad (2\text{-}35)$$

其中，C_{Miss} 和 $C_{FalseAlarm}$ 分别表示 FRR 和 FAR 的权重，P_{Target} 表示目标说话人的先验概率，P_{Miss} 和 $P_{FalseAlarm}$ 表示 FRR 和 FAR。在 NIST SRE 2002[5]中，C_{Miss}、$C_{FalseAlarm}$ 和 P_{Target} 的取值分别为 10、1 和 0.01，本文实验中计算 *mDCF* 时，使用相同的数值。

表 2.1 参考说话人模型数量选择实验

K	**EER(%)**			***mDCF*($\times 10^{-2}$)**		
	1(秒)	**2(秒)**	**5(秒)**	**1(秒)**	**2(秒)**	**5(秒)**
64	21.5	18.6	14.6	9.55	6.52	5.91
128	19.8	17.1	13.2	9.37	6.39	5.64
256	18.7	16.2	11.3	9.04	6.09	5.31
512	18.9	16.3	11.2	9.15	6.11	5.33

表 2.1 中的 *K* 表示参考说话人的数量，参考说话人模型的混合数根据经验值选择了 1,024，参考说话人的数量太少使得每个距离度量精度不足，而参考说话人的数量太多会由于数据不足使得某些参考说话人的代表性不好，在本章后面的实验中，根据表 2.1 中的实验结果选择 $K = 256$。

实验二　挑选与语音距离较小的 RSM 度量相似程度

参考说话人需要有足够的声学空间覆盖范围和精度，因此参考说话人的数量要足够多以保证每窗语音都能够找到比较相似的参考说话人模型。而由于多个参考说话人模型覆盖的声学空间较广，而一般来说一窗语音仅与一小部分参考说话人模型的距离较小而与大部分参考说话人模型的距离较大，距离较大的参考说话人模型对于窗内语音的区分能力相对较弱，在计算语音之间的相似程度时会带来一定程度的精度影响，因此对距离计算进行了修改，比较两段语音在距离较小的说话人模型上的距离，将与语音距离较大的参考说话人模型去除，使用距离较小的参考说话人模型来进行距离度量。根据表 2.2 中的实验结果选择了距离最小的前 64 个参考说话人模型度量距离。

表 2.2 挑选与语音距离较小的 RSM 度量相似程度

K-top	**EER(%)**			***mDCF*($\times 10^{-2}$)**		
	1(秒)	**2(秒)**	**5(秒)**	**1(秒)**	**2(秒)**	**5(秒)**
16	18.6	16.3	11.7	8.93	6.25	5.37
32	18.5	15.9	11.3	8.89	6.12	5.30
64	18.3	15.7	11.1	8.66	6.00	5.24
128	18.4	16.1	11.3	8.78	6.11	5.28
256	18.7	16.2	11.3	9.04	6.09	5.31

实验三　短语音说话人确认性能比较

在基于距离度量的说话人分割算法中，其分割性能与距离度量的性能正相关。一般来说，一种好的距离度量准则应该既能很好地反映出不同说话人之间的差异，又能使得同一说话人自身语音之间的差异较小。对于两段语音，如果两段语音属于同一说话人则根据距离度量准则计算的二者之间的距离较小；如果两段语音属于不同的说话人则两段语音之间的距离较大。本章中使用一个说话人确认系统来比较不同距离度量准则的度量精度，不训练说话人模型，每次说话人确认直接使用两段较短的语音使用距离度量准则计算其距离，最后分析说话人确认系统的性能来评价距离度量的性能。如果基于某种距离度量准则的说话人确认系统性能好，则其距离度量就更准确、更能区

分不同的说话人，相应的分割性能也会随之好。因此，本文比较了 BIC，GLR 和 RSM 三种算法在不同语音长度下的确认性能。

表 2.3 三种度量准则的说话人确认性能对比

算法	EER(%)			*mDCF*($\times 10^{-2}$)		
	1(秒)	2(秒)	5(秒)	1(秒)	2(秒)	5(秒)
BIC	22.6	19.3	13.1	9.24	8.61	6.97
GLR	23.1	20.0	12.9	9.74	9.02	6.77
RSM	18.3	15.7	11.1	8.66	6.00	5.24

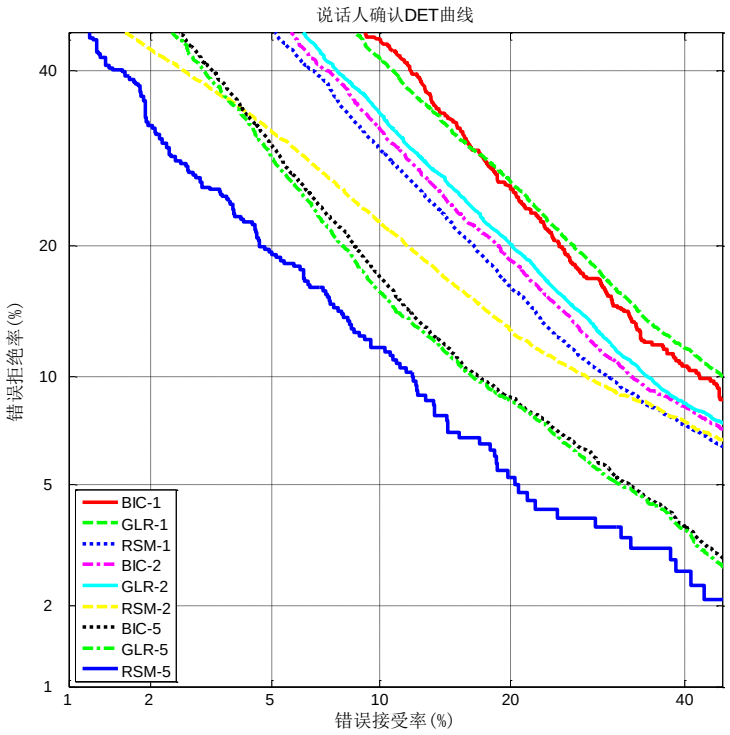


图 2.5 短语音说话人确认性能比较实验

在三种长度的数据集合上，使用 RSM 的确认系统的性能较使用 BIC 或 GLR 的确认系统均有不同幅度的提高，也就说明了基于 RSM 的距离度量准则精度比后二者要好。在语音长度分别为 1 秒、2 秒和 5 秒的条件下，与 BIC 相比，EER 分别相对下降 19.0%、18.7%和 15.3%，*mDCF* 分别相对下降 6.3%、30.3%和 24.8%。

实验四　窗宽选择实验

不同的窗宽对于分割性能存在一定的影响，如果说话人一次发音的平均时间较长，则窗宽较长的性能会好些，因为数据多距离度量的准确性和稳定性都好；如果说话人一次发音的平均时间较短，则窗宽长些性能会差，因为窗内可能存在多个说话人转换导致算法分割性能变差。

本文在统计分割结果时，给定了一个 0.3 秒的容错范围，也就是说如果得到的某分割点与实际分割点的距离小于 0.3 秒，则认为该分割点是正确的分割点，否则判为误警点或漏检点。根据误差容错范围，窗移选择为 0.3 秒。

由式（2−34）可得，阈值与 α 取值正相关，α 的大小决定着分割点判断的松紧程度，α 取值变小则 MDR 会随之变小而 FAR 会随之变大；α 取值大则 FAR 会随之变小而 MDR 会随之变大。尽管 MDR 越小也好，但通过降低 α 的取值使 MDR 变小会导致 FAR 过大，使得分割后的语音段的平均长度太短，不利于后面的聚类和说话人辨认，实验中根据两种错误率之和最低的原则来选择 α 的取值。

表 2.4 窗宽选择实验−1

数据库	窗宽（秒）	FAR(%)	MDR(%)
BNEWS	1.0	25.7	10.8
	1.5	24.2	11.3
	2.0	**22.5**	**11.6**
	2.5	21.9	12.5

表 2.5 窗宽选择实验−2

数据库	窗宽（秒）	FAR(%)	MDR(%)
SWBD	0.8	26.8	23.6
	1.0	**25.1**	**24.0**
	1.5	24.7	24.9
	2.0	23.6	26.2

从表 2.4 和 2.5 中可以看出，在 BNEWS 和 SWBD 两个数据库上，随着窗宽的变长 MDR 会变大而 FAR 变小，这是因为窗宽较长会使短发音被漏检，同时由于一窗内数据的增多使得距离度量的精度和稳定性得到提高，从而使得 FAR 降低。在后续的比较实验中，BNEWS 数据库上的窗宽选择为 2 秒，SWBD 数据库上的窗选择为 1 秒。

实验五　性别信息和波峰波谷信息应用实验

表 2.6 和 2.7 中，RSM 表示未使用性别信息的多参考说话人模型；G 表示使用了性别信息；PV 表示使用了波峰波谷信息优化；对 RSM 加入性别信息之后，尽管幅度不大但对分割性能仍有改善。应用 PV 要求两个分割点之间的语音长度达到一定长度，因为如果较短的语音段再取一半进行距离计算会因为数据少而导致精度较差，PV 对 BNEWS 数据效果较好而对 SWBD 效果一般。

表 2.6 应用性别信息和波峰波谷信息分割实验−1

数据库	算法	FAR(%)	MDR(%)
BNEWS	RSM	22.5	11.6
	RSM+G	22.6	11.4
	RSM+PV	22.1	11.0
	RSM+G+PV	21.5	10.7

表 2.7 应用性别信息和波峰波谷信息分割实验−2

数据库	算法	FAR(%)	MDR(%)
SWBD	RSM	25.1	25.0
	RSM+G	24.6	24.8
	RSM+PV	24.5	24.6
	RSM+G+PV	24.2	24.5

实验六　分割性能比较实验

本文比较了基于 BIC、GLR、DISTBIC 和 RSM 的四种说话人分割算法的性能，结果参见表 2.8、表 2.9。使用相对错误率下降（Error Reduction Rate，ERR）来描述性能改变程度，ERR 的定义如下：

$$ERR = \frac{E_{old} - E_{new}}{E_{old}} \times 100\% \qquad (2\text{-}36)$$

其中，E_{new} 是新算法的错误率，E_{old} 是旧算法的错误率。

表 2.8 BIC、GLR、DISTBIC 和 RSM 分割性能比较-1

数据库	算法	FAR(%)	MDR(%)
BNEWS	BIC	23.5	15.8
	GLR	24.3	16.2
	DISTBIC	23.2	16.4
	RSM+G	22.1	11.0
	RSM +G+PV	21.5	10.7

表 2.9 BIC、GLR、DISTBIC 和 RSM 分割性能比较-2

数据库	算法	FAR(%)	MDR(%)
SWBD	BIC	28.8	28.1
	GLR	30.2	26.3
	DISTBIC	27.9	26.5
	RSM+G	24.5	24.6
	RSM +G+PV	24.2	24.5

BIC 算法基于式（2-4）、GLR 算法基于式（2-7），DISTBIC 算法预分割使用基于式（2-7）的 GLR 距离和基于式（2-10）的 KL 距离，优化使用基于式（2-4）BIC 算法。在 BNEWS 数据库上四种算法使用的窗长都是 2

秒，窗移为 0.3 秒；在 SWBD 数据库上四种算法使用的窗长都是 1 秒，窗移为 0.3 秒。

从表 2.8 和表 2.9 可以看出，基于 RSM 的分割算法取得了较好的分割性能，与传统的 DISTBIC 算法相比，在 BNEWS 数据库上漏检率相对下降 *ERR*=34.8%，总错误率（FAR＋MDR）相对下降了 *ERR*=18.7%，在 SWBD 数据库上在电话交谈语音库 SWBD 上漏检率相对下降了 *ERR*=7.5%，总错误率相对下降 *ERR*=10.4%。由于多参考说话人模型能够较好地覆盖声学空间代表说话人的发音共性，在距离计算中利用了多参考说话人模型作为先验知识后，使得对相邻两窗语音的说话人特性描述更准确，从而使计算得到的“距离”能够比较好的反映出不同说话人的发音差异。

2.4.3　讨论

从上面的实验可以看出，RSM 能够较好的改善较短窗长条件下两窗语音之间距离度量的准确性，在新闻采访语音和电话交谈语音两种说话人分割数据库下，相对于基线系统性能均有明显提升。但是电话交谈语音的分割性能与新闻采访语音相比性能明显下降，这是因为电话交谈语音中存在很多的较短发音，以中文电话交谈为例，会有较多的一两汉字的短发音，如“是、行、嗯、你好”等，此时发音的平均长度在 0.5 秒左右，这样的说话人转换点很难被检测到，导致漏检情况增多。

2.5　小结

本章重点分析了影响基于距离度量的说话人分割算法性能的主要因素，现有的距离度量准则在窗宽较短的条件下对窗内语音的说话人发音特性的描述不够准确，从而导致距离度量容易产生偏差与不稳定。在此基础上，提出了一种基于参考说话人模型的说话人分割聚类算法。该算法包括四个步骤：获取参考说话人模型、分窗计算距离、分割和优化。

1．采用参考说话人模型描述声学空间，使用语音与多个参考说话人的距离来描述说话人的发音特性；并将两窗语音的发音特性之间的距离作为说话人分割的一种度量准则。实验结果表明，基于参考说话人模型的度量准则较 GLR、BIC 方法能够更准确的度量较短的两段语音之间的距离，在新闻采访语音数据库和电话交谈语音数据库上分割效果均有一定提高。应用波峰波谷信息进行分割优化后，性能得到进一步优化。

2．加入性别信息帮助分割。使用区分性强的信息对说话人分割会有帮助，性

别信息是能够区分是否发生说话人转换的十分有效的信息，因此加入性别信息参考说话人模型后，算法的分割性能得到进一步改善，可在未来进一步开展区分性特征的应用研究。

3．对于新闻采访语音，算法分割性能的提升幅度明显，在说话人一次发音平均时间较短的电话交谈语音上，分割性能有改善但幅度较新闻采访语音有明显的下降。因为电话交谈语音中存在很多的较短发音，这些短发音很难被检测到，导致漏检情况增多。

第3章　基于音素识别和文本相关的说话人分割算法

基于参考说话人模型的说话人分割聚类算法较 BIC 和 GLR 距离度量准则在新闻采访语音上有了一定改善，但在电话交谈语音上的性能较新闻采访语音还存在较大差距。电话语音的特点，就是既有较长的单人语音段又有较短的单人语音段，其中较短的单人语音段对分割性能的影响很大。在第 2 章中分析过，说话人分割算法的性能在一定程度上依赖于较短的两段语音之间距离度量的准确性，BIC、GLR 和 RSM 准则都只使用了语音的低层特征而没有利用高层特征或专家知识，能否借助于高层特征或专家知识来进行分割，本章中进行了一定的探索。

本章的内容安排如下：3.1 介绍一些利用高层特征的说话人分割算法；3.2 介绍基于音素识别和文本相关的说话人分割算法；3.3 给出实验结果和分析；3.4 是本章的小结。

3.1　利用高层信息的说话人分割算法介绍

近年来，逐渐有研究人员利用高层信息和专家知识进行说话人分割，如利用了不同频带携带的说话人区分性信息不同的 FREQDIST 算法[107]，融合长时特征和低层特征的算法[55,56]，下面简单介绍一下这两种算法。

3.1.1　基于区分性频域特征的说话人分割算法

Boehm 和 Pernkopf 2009 提出了 FREQDIST 算法[107]，利用语音的发音基理等知识分析了频谱中不同频带的说话人区分能力，使用归一化技术对频谱进行了重新加权，区分性较弱的频谱信息使用基于熵的方法进行了去除，使用区分性强的频谱信息之间的欧式距离作为距离度量，其基本思想和主要过程如下：

（1）随着说话人声道的变化语音的共振峰会产生变化。第 1、2 有时甚至第 3 共振峰对不同的说话人不会有太多变化，它们发生变化主要依赖于是发音的内容；前 2 到 3 个共振峰对于语音识别作用较好，但对于说话人分割来说可能并不合适。

（2）说话人分割中的常用特征如 MFCC、LPCC 等使用了所有的可利用频率成分，这其中包含了较难区分说话人的共振峰，如前 2 到 3 个共振峰。

（3）第 4（有时是第 3）和更高的共振峰是更加说话人相关的，在不同的说话人之间的有更强的变化。

（4）利用归一化技术进行频谱加权，去除频率和能量之间的关系，并将每一帧语音的能量和归整为 1；利用熵来判断一帧语音的区分性的大小，熵越大则区分性越小，区分性小的语音帧在计算窗间的距离时不被使用。

（5）分窗并使用欧式距离来度量两窗语音之间的距离。

在文献[108]的实验中该算法与 DISTBIC 算法相比，FAR 从 10.4%下降到 5.1%，MDR 从 13.1%下降到 7.0%，实验数据的说话人平均发音长度达到的 7.5 秒，分割点的误差允许范围为 1 秒，该算法要求语音的采样频率较高，16,000Hz 或更高。

3.1.2　短时特征与长时特征融合的说话人分割算法

Friedland 等 2009 提出了融合使用短时特征和长时特征的说话人分割算法[55,56]，首先挑选了 70 种不同的长时特征，从中选择说话人区分性强的 10 种特征与短时特征融合，该算法在 NIST RT 07[6]测试集合上取得了较好的性能。

长时特征选择了基音周期（Pitch）、响度（Loudness 或 Energy）、共振峰（Formants）、谐波噪音比（Harmonics to Noise Ratio，HNR）和长时平均频谱（Long Term Average Spectrum, LTAS）作为基本特征。然后对这些基本特征进行各种变化，如取其最小值、最大值、中值、均值、差分和标准差等，其中共振峰选择了第 1 到第 5 共振峰及它们的各种变化值。

根据费舍尔准则判断各种特征对说话人区分性的强弱，选择区分性最强的 10 种特征与短时特征融合。在特征层面进行了融合，短时特征使用了 19 维的 MFCC 和 10 维的长时特征融合成 29 维的特征，计算距离时对两种特征进行了加权处理。

说话人分割时使用了融合分割算法，基于 BIC 距离度量的分割算法与基于 HMM 的模型搜索算法相融合。

3.2　基于音素识别和文本相关的说话人分割算法

3.2.1　基本思想

上一小节中介绍的两种利用高层信息或专家知识的分割算法在一定条件下都取得了较好的分割效果，为改善分割性能提供了一种新的思路。说话

人分割问题可在一定程度上转换为两段较短语音之间的相似性度量问题，而文本相关的说话人识别在度量两段短语音之间相似性的精度较好，语音长度可以短到仅有一两个字，但由于是两段语音在内容上的高度一致使得文本相关的距离度量准确性相当高。本文中提出了一种利用音素信息的、基于音素识别和文本相关说话人识别的说话人分割算法，借助音素识别技术对语音进行切分得到音素序列，分窗后对两窗内相同的音素进行文本相关的相似性度量并将其结果作为两窗语音的相似性度量结果。图 3.1 是算法的示意图。

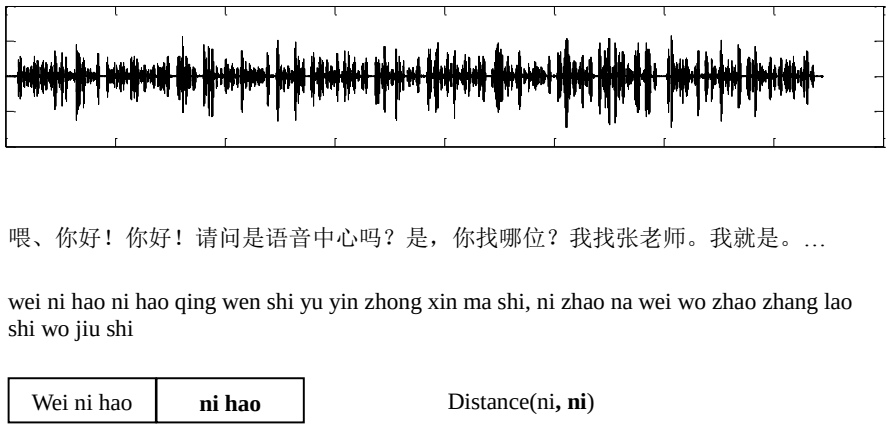


图 3.1 基于音素识别和文本相关的说话人分割算法示意图

图 3.1 中，一段语音经过音素识别后得到音素序列，分窗后对两窗内相同的音素内容进行文本相关的说话人识别，将识别结果作为度量距离进行说话人分割。

3.2.2 算法描述

文本相关的说话人识别根据两段相同内容的语音之间的距离来判断两段语音是否属于同一个说话人。假设相邻的 L 和 R 两窗语音内都有音素 P 的发音，分别为 P_L 和 P_R，如果两窗语音属于同一说话人，P_L 和 P_R 之间的差异一般较小，使用文本相关的说话人识别 P_L 和 P_R 得到的距离也会较小；如果两窗语音属于不同的说话人，P_L 和 P_R 之间的差异一般较大，使用文本相关的说话人识别 P_L 和 P_R 得到的距离也会较大。基于音素识别和文本相关的说话人分割算法包括五个步骤：音素识别、分窗与文本相关识别、加权处理、无相同音素处理和分割点判断。分割点判断同第 2 章中的算法相同，下面详细介绍算法的其余步骤。

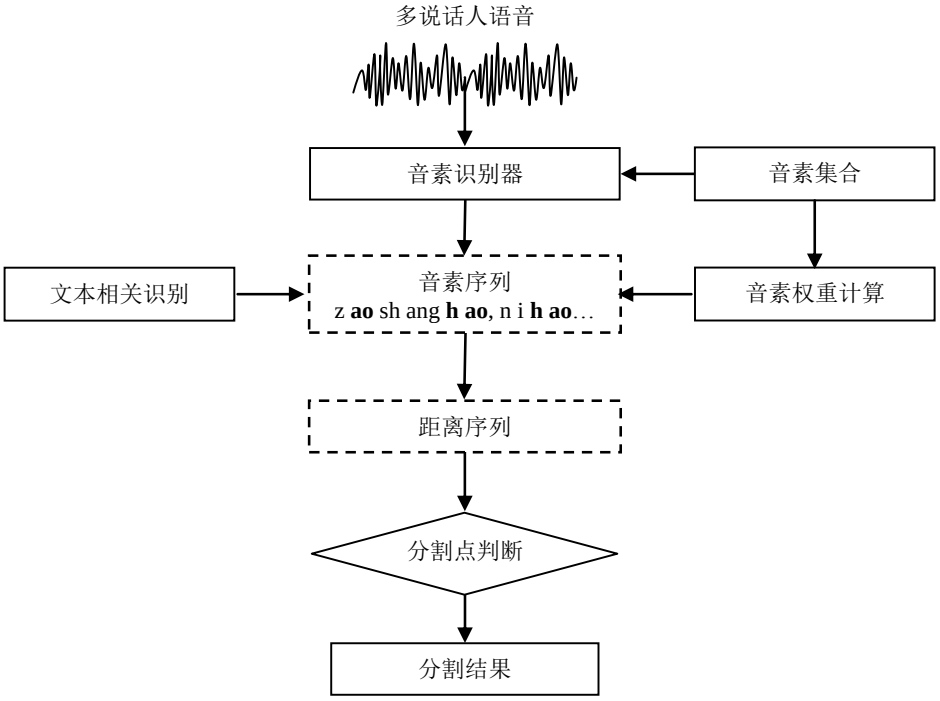


图 3.2 基于音素识别和文本相关的说话人分割算法流程

3.2.2.1 音素识别

音素识别就是使用音素识别器对输入语音进行音素识别以获得音素序列。只有可靠的音素标注才能进行保证算法的性能，因此对音素识别器的选择很重要。在目前已报道的音素识别引擎中，BUT（Brno University of Technology，BUT）的基于神经网络和维特比解码的公开的音素识别器，识别率高性能最好[108]。因此本文选用 BUT 提供的英语识别引擎，对输入语音解码以得到音素序列。

3.2.2.2 分窗与文本相关识别

分窗的过程与第 2 章第 2.3 小节中的分窗方法相同，不同之处在于本章是对音素序列分窗，而音素是有一定时间宽度的，在本章中，如果窗移动后如果左边分析窗起点在一个音素的中间位置则将其左移到该音素的起点时刻，右边分析窗如果在一个音素的中间位置则将其右移到该音素的终点时刻。

文本相关的说话人识别使用动态时间规整（Dynamic Time Warping，DTW）[110-113]算法度量相邻两窗内相同音素之间的距离，DTW 采用动态时间

规划方法以将测试语音弹性缩放，使得测试语音的长度能够对齐训练语音，得到测试特征序列和训练特征序列的匹配距离，最后得到匹配分数。DTW 中的距离度量方式使用似然分数差，首先挑选训练特征在 UBM 上的核心分布并计算似然分，测试特征在 UBM 核心分布上计算似然分，二者差的绝对值即为训练特征和测试特征之间的距离，距离越小表示二者越接近。假设 S_1 和 S_2 分别是相邻两窗语音的音素序列：

$$S_1:\ P_{11},...,P_{1m},...P_{1K_1},m\in[1,K_1]$$

$$S_2:P_{21},...,P_{2n},...P_{2K_2},n\in[1,K_2]$$

其中，K_1 和 K_2 分别是音素序列 S_1 和 S_2 中的音素个数。音素序列 S_1 和 S_2 之间的文本相关的距离：

$$D(S_1,S_2)=\frac{1}{N_v}\sum_{i=1}^{N}\alpha_i d_V\left(P_{S_1}^i,P_{S_2}^i\right) \tag{3-1}$$

其中，$d_V\left(P_{S_1}^i,P_{S_2}^i\right)$ 是 S_1 和 S_2 中的音素 P^i 之间的距离度量，α_i 是音素 P^i 的距离权重，$P_{S_1}^i$ 和 $P_{S_2}^i$ 分别表示音素序列 S_1 和 S_2 中音素 P^i，N_V 是音素集合中同时在 S_1 和 S_2 中出现的音素的数量。

如果音素序列 S_1 和 S_2 中都不存在音素 P^i，则 $d_V\left(P_{S_1}^i,P_{S_2}^i\right)=0$；如果 S_1 和 S_2 中都存在音素 P^i，则：

$$d_V\left(P_{S_1}^i,P_{S_2}^i\right)=\frac{1}{K_1^iK_2^i}\sum_{j=1}^{K_1^i}\sum_{k=1}^{K_2^i}DTW\left(P_{S_1}^{ij},P_{S_2}^{ik}\right) \tag{3-2}$$

其中，$DTW\left(P_{S_1}^{ij},P_{S_2}^{ik}\right)$ 是音素序列 S_1 中的第 j 个音素 P^i 和 S_2 中第 k 个音素 P^i 之间的 DTW 距离，K_1^i 和 K_2^i 分别是音素序列 S_1 和 S_2 中音素 P^i 的数量。

3.2.2.3　音素加权处理

不同音素由于其发音方式和发音内容不同，对于说话人的区分性会存在着一定差异，如辅音与元音的说话人区分性不相同，不同元音之间也会存在一定的差异。因此，不同音素的距离采用不同的权重可能会比等权重对于说话人的区分性会更好。

本文中利用 F-Ratio[97]来评价不同音素的说话人区分性，利用音素的 F-Ratio 值进行归一化来调整音素权重。F-Ratio 利用音素在说话人之间的变化 b 与音素在同一说话人不同次发音之间的变化 w 来计算，一个音素的 b 与 w 的比值越大，说明该音素在不同说话人之间差异大而在同一个说话人的多次发音之间差异小，该音素的区分性就越强；反之如果 b 与 w 的比值越小，则该音素的区分性就越弱。音素 P^i 的 F-Ratio 值为：

$$\overline{d_i} = AVG\left(DTW\left(P_j^{im}, P_k^{in}\right)\right),\ j \neq k,\ j,k \in [1,G] \tag{3-3}$$

式（3-3）中，G 代表开发集中说话人的数量，P_j^{im} 和 P_k^{in} 分别是说话人 j 的音素 P^i 的第 m 次发音和说话人 k 的音素 P^i 的第 n 次发音，$\overline{d_i}$ 是开发集中任意两个分别被不同说话人发出的音素 i 之间 DTW 距离的平均值，AVG 表示求平均。

$$b_i = \frac{1}{G(G-1)}\sum_{m=1}^{K_{ji}}\sum_{n=1}^{K_{ki}}\left(DTW\left(P_j^{im}, P_k^{in}\right) - \overline{d_i}\right)^2,\ j \neq k,\ j,k \in [1,G] \tag{3-4}$$

式（3-4）中，b_i 描述了音素 i 在说话人之间的变化情况，其中，j 和 k 代表开发集中的任意两个不同的说话人，K_{ji} 和 K_{ki} 分别是说话人 j 和 k 的语音中音素 P^i 的发音次数。

$$\overline{d_{is}} = \frac{1}{C_s\left(C_s - 1\right)}\sum DTW\left(P_s^{ij}, P_s^{ik}\right),\ j \neq k,\ j,k \in [1,C_s] \tag{3-5}$$

式（3-5）中，$\overline{d_{is}}$ 是同一个说话人 s 语音中音素 P^i 的不同次发音之间 DTW 距离的平均值，C_s 是说话人 s 语音中音素 P^i 的发音次数。

$$w_i = \frac{1}{G}\sum_{s=1}^{G}\frac{1}{C_s\left(C_s - 1\right)}\sum_{j=1}^{C_s}\sum_{k=1}^{C_s}\left(DTW\left(P_s^{ij}, P_s^{ik}\right) - \overline{d_{is}}\right)^2,\ j \neq k \tag{3-6}$$

式（3-6）中，w_i 是音素 P^i 在说话人自身不同次发音之间的变化情况。音素 P^i 的 F-Ratio 值为：

$$bw\left(P_i\right) = \frac{b_i}{w_i} \tag{3-7}$$

使用式（3-8）来调整音素 P^i 的距离权重。

$$\alpha_i == \frac{bw(P_i)}{\sum_{j=1}^{N} bw(P_j)} \tag{3-8}$$

其中，α_i 是音素 P^i 的新的距离权重。

不同的音素，其音素距离的值域范围可能变化很大，有的音素的距离值域范围较大而有的音素则较小。假设有两个音素 P^A 和 P^B，音素 P^A 值域范围较大但其说话人区分性较小，音素 P^B 值域范围较小但其说话人区分性较强，按照式（3-1）计算得到的距离值就很可能更大程度上依赖于音素 P^A 的距离值，而理想情况是最终的距离值更大程度上依赖与说话人区分性强的音素的距离值，因此对 DTW 距离值进行归一化处理，以避免值域范围大的音素的距离淹没值域范围小的音素的距离。本文中使用式（3-9）进行归一化处理。

$$d_V\left(P_{S_1}^i, P_{S_2}^i\right)_{norm} = \frac{1}{1+e^{-d_V\left(P_{S_1}^i, P_{S_2}^i\right)}} \tag{3-9}$$

3.2.2.4　相邻窗内无相同音素情况的处理

有一种特殊情况是必须考虑的，就是音素集合中任何一个音素在相邻的两窗中都没有同时出现，这种情况比较少，图 3.3 是相邻两窗不存在相同音素的窗数与总窗数的比例情况。这种情况下，采用了两种方法来解决，一是适当增加窗宽，即左窗向左扩展，右窗向右扩展。二是线性插值方法。使用其左右相邻的距离值以及两侧距离值的变化方向来估计该点的距离值。本文后面的实验中先进行窗宽扩宽，如果还不存在相同音素则利用插值法获取该点的距离值。

3.3　实验结果与分析

3.3.1　实验数据和设置

考虑对比实验的需要，实验中数据最好是带有音素标注信息的，因此选择了 TIMIT 语音数据库[113]，TIMIT 数据库共包含 630 个说话人，每个说话人有 10 段语音，每段语音大约 3 秒长，对语音进行分割拼接后得到分割测试数据，说话人每次发音的平均长度参见表 3.1。

分割点判断方法与第 2 章中的分割点判断方法相同。在统计分割结果 FAR 和 MDR 时，与第 2 章一样给定了一个 0.3 秒的容错范围，窗移设定为 0.3 秒，根据对相邻两窗中是否存在相同音素的分析，窗宽选择为 2 秒。

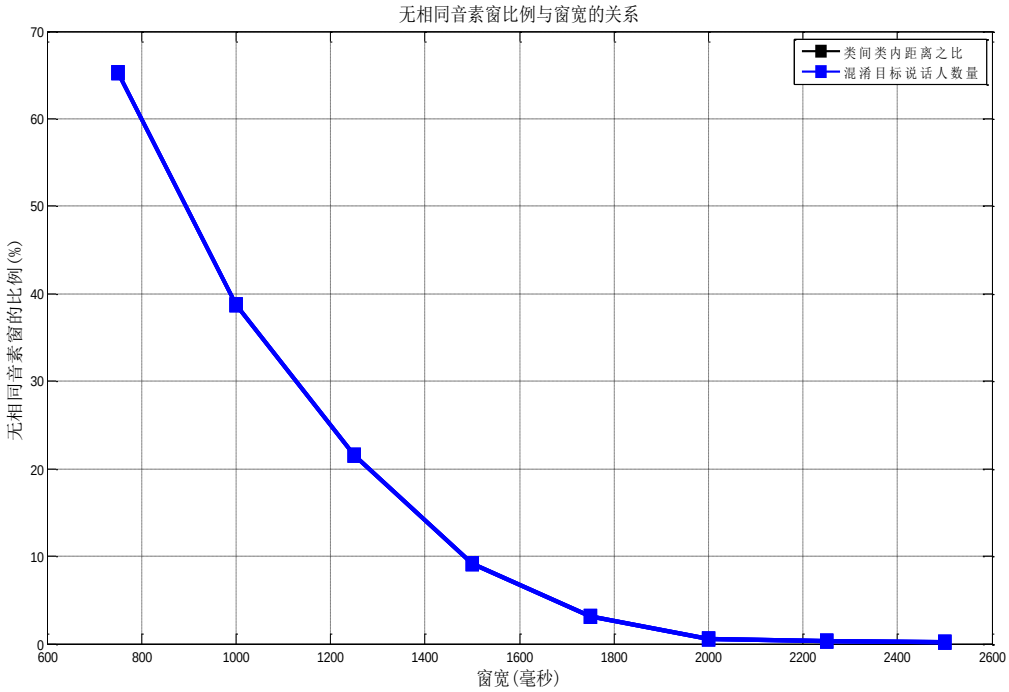


图 3.3 相邻两窗不存在相同音素的窗数与总窗数之比

表 3.1 每次发音长度情况

平均发音长度（秒）	语音段数	百分比(%)
1~2	1,195	29.1
2~6	2,588	63.2
6~10	317	7.7

3.3.2　实验结果与分析

实验一　短语音说话人确认性能比较

同基于 RSM 的距离度量类似，设计了一个说话人确认实验来分析几种距离度量方法 BIC，GLR、RSM 和 PR 的说话人确认性能。PR 代表基于音素识别和文本相关的距离度量方法。因此，本文在比较分割性能之前，先比

较了三种算法的说话人确认性能。每次确认的两段语音中都包含有相同的音素但发音内容并不相同，实验数据是 TIMIT 数据根据标注进行一定的挑选利用分割拼接得到，语音长度有 3 种：1 秒、2 秒和 5 秒。

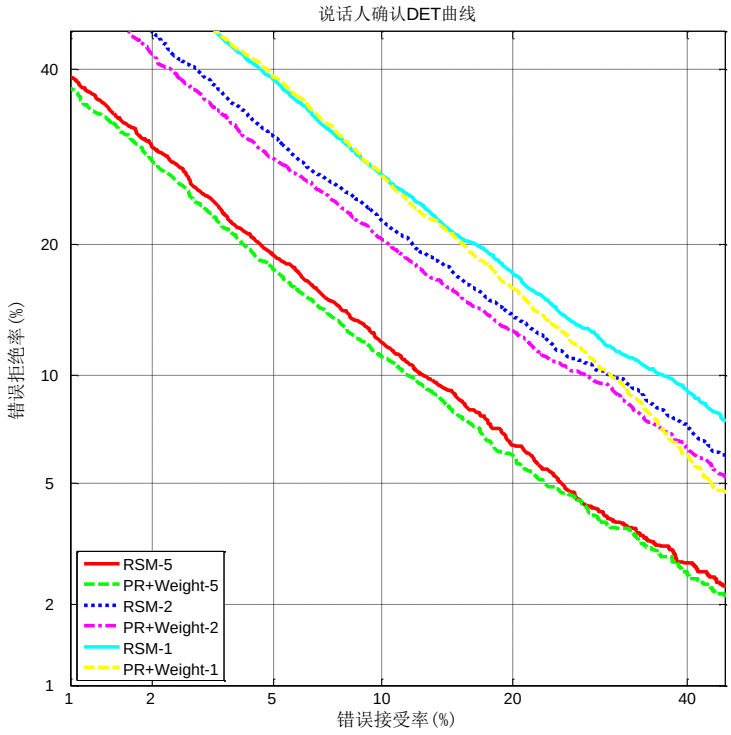


图 3.4 说话人确认 DET 曲线

表 3.2 短语音说话人确认性能比较实验

算法	EER(%)			*mDCF*($\times 10^{-2}$)		
	1(秒)	2(秒)	5(秒)	1(秒)	2(秒)	5(秒)
BIC	22.4	19.2	13.2	7.64	7.25	5.76
GLR	23.2	19.9	12.8	7.70	7.38	5.37
RSM	18.6	16.3	11.2	6.94	6.32	4.85
PR	18.2	16.0	11.0	7.26	6.11	4.70
PR+Weight	18.0	15.5	10.8	7.18	6.08	4.66

在三种长度的数据集合上，基于音素识别和文本相关的距离度量方法的说话人确认相同的性能较使用 BIC、GLR 和 RSM 的说话人确认系统均有不

同程度的提高。在语音长度分别为 1 秒、2 秒和 5 秒的条件下，与基于 RSM 的方法相比，EER 分别相对下降了 3.2%、4.9%和 4.3%。PR+Weight 表示对音素权重根据音素 F-Ratio 值进行了调整，进行音素权重调整之后，PR 方法的说话人确认性能得到了进一步提高。

实验二　分割性能比较

基于音素识别和文本相关的说话人分割方法首先使用音素识别工具对语音进行识别，其分割性能对音素识别器性能的依赖性如何，本文设计了一个实验进行分析。实验是使用已经做好标注的对话语音，一是已知准确的音素序列下的极限分割性能，使用标注信息作为音素序列进行分割；二是使用 BUT 音素识别器进行音素识别，使用其识别结果进行分割。并同 BIC、GLR 和 RSM 进行了对比。

表 3.3 分割实验结果

算法	FAR(%)	MDR(%)
BIC	28.2	28.3
GLR	31.5	25.9
RSM	25.4	24.8
PR	25.3	21.5
Trans	22.1	20.3

表 3.3 中 PR 算法表示基于音素识别和文本相关的分割算法，Trans 表示使用语音的音素标注作为音素序列的分割算法。从表 3.3 中可以看出，在实验条件下，PR 算法较 BIC、GLR 和 RSM 均有改善，与 RSM 算法相比，漏检率相对下降 15.4%，总错误率相对下降 6.8%；与 GLR 算法相比，漏检率相对下降 16.9%，总错误率相对下降 18.5%；与使用标注信息相比，漏检率增加 1.2%，总错误率增加了 4.4%，尽管音素分割存在一定错误，但分割结果中的“相同音素”在一致性或相关性还是得到了增强，因此其距离度量精度得到了提高，使分割性能得到改善。

3.4　小结

本章重点讨论了如何借助高层信息或专家知识来进一步提高较短的两窗语音之间距离度量精度的方法，提出了一种基于音素识别和文本相关的说话人分割算法，借助音素识别技术获取音素信息，分窗后对两窗内相同的音素进行文本相关的相似性度量并将其结果作为两窗语音的相似性度量结果，该算法在限定条件下取得了一定的分割性能的改善，与 RSM 算法相比，漏检率相对下降 15.4%，总错误率相对下降 6.8%；与 GLR 算法相比，漏检率相对下降 16.9%，总错误率相对下降 18.5%。但 PR 算法的复杂度高，受音素识别器性能以及语音内容的限制。

第4章　基于类纯度约束的说话人聚类算法

由于分割点中误警较多或说话人平均发音长度较短的原因，说话人分割之后得到的单说话人语音段的长度可能较短，直接用来进行说话人辨认会因为数据少而影响说话人辨认的性能，因此需要对分割结果进行说话人聚类处理，将属于同一个说话人的语音聚成一类从而增大单说话人语音段长度。说话人聚类的理想结果是：聚类数等于说话人数，每一类只包含一个说话人的语音。而大多数情况下，多说话人语音中说话人的数目和身份信息都是未知的，其中说话人每次发音的长度也变化很大，有的时间较长也有的较短。这些因素都会影响到说话人聚类的效果，也是本章的重点研究内容。

本章内容安排如下：4.1 简单介绍一些常用的说话人聚类算法并给出相应的分析；4.2 介绍说话人聚类算法的评价指标；4.3 介绍本文提出的基于类纯度约束的说话人聚类算法；4.4 给出实验结果和分析；4.5 是本章的小结。

4.1　常用聚类算法

说话人聚类（Speaker Clustering）[115-117]的目的就是从一个数据集中找出所有属于同一说话人的数据。说话人聚类技术根据聚类任务的不同，可分为有监督聚类和无监督聚类两种。前者在聚类时事先知道说话人数目，而在后者中说话人数目则是未知的。K 均值聚类（K-Means Clustering）[98,99]算法是有监督聚类方法中最具有代表性的方法，分层凝聚聚类（Hierarchical Agglomerative Clustering，HAC）算法[17]和基于隐马尔可夫模型（Hidden Markov Model，HMM）[29,118]的说话人聚类算法则是无监督聚类方法中最常用的两种。

由于在大规模目标说话人检测中无法事先获得多说话人语音中说话人的数目，因此只能使用无监督的聚类方法。下面先简单介绍一下常用的两种无监督说话人聚类算法。

4.1.1　HAC算法

HAC 算法是最常用的无监督说话人聚类中算法，Gish 等 1991[17]、Jin 等 1997[115]、 Solomonoff 等 1998[116]、Chen 和 Gopalakrishnan 1998[13]、Reynolds 等 1998[25]、Johnson 和 Woodland 1998[118]、Faltlhauser 和 Ruske

2001[119]、Moh 等 2003[120]、Liu 等 2005[121]、王炜等 2008[122]都进行了研究，对距离度量准则和聚类停止条件等进行了研究。HAC 算法的流程示意参见图 4.1。

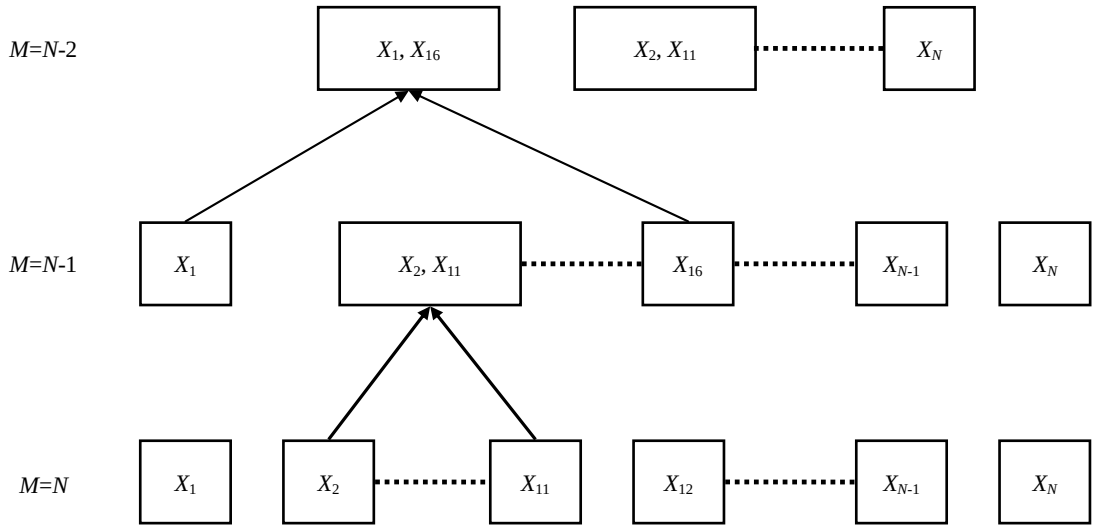


图 4.1 HAC 算法示意图

图 4.1 中$\{X_1,X_2,\ldots,X_N\}$是一组语音段，N 是聚类初始时的语音段数，M 是聚类数，聚类开始时 $M=N$，按照选定的距离度量准则计算任意两段语音之间的距离并将距离最近的两段语音合并，聚类数 M 也随之变为 N-1，重复上述步骤直到满足收敛条件。收敛条件一般是距离最小的两段语音之间的距离大于给定的阈值。常用的距离度量准则有 GLR、BIC、KL 距离以及 CLR 等。

HAC 算法的优点是概念简单清晰、算法复杂度低；缺点是聚类结果对阈值较为依赖，对其变化非常敏感，很难设定普适的阈值。

4.1.2　基于HMM的聚类算法

基于 HMM 的聚类算法一般都和说话人分割融合进行，HMM 中的每个状态代表一个类也即一个说话人，状态之间的转移代表说话人的转换。Meignier 等 2002[29]，Ajmera 等 2002[117]都进行了研究。基于 HMM 的说话人聚类算法中需要进行说话人分割，待聚类语音看作是马尔可夫链的一组观测值，初始时 HMM 的状态数等于语音段数[117]，主要过程如下：

（1）HMM 中的每个状态的概率密度函数（Probability Density Function，PDF）使用一个高斯混合模型表示，使用 EM 算法将每段语音训练成一个高斯混合模型。

（2）以似然比作为距离度量，将对应模型之间距离最近的两段语音合并，使用 EM 算法重训新语音段的高斯混合模型，且新模型的混合数更大。

（3）使用步骤 2 中生成的新模型和 Viterbi 解码算法重新对整段语音进行说话人分割，如果 Viterbi 解码之后路径得分增加，则说明合并有效；如果 Viterbi 解码之后路径得分降低，则说话人合并无效，停止合并。

基于 HMM 的聚类算法优点是性能好；缺点是算法复杂度高，需使用 Viterbi 解码算法和 EM 算法，而且要迭代多次，不断进行重分割和模型训练，时间消耗大，在对速度有严格要求的说话人检测中很难应用。

4.2　说话人聚类算法的评测指标

说话人聚类的评价指标有类纯度、说话人纯度等[117]。类纯度衡量同一类中的语音来自同一说话人的集中程度，说话人纯度衡量的是同一说话人的语音数据的分散程度。

类纯度和说话人纯度的定义为[117]：

$$\overline{p} = \frac{1}{N}\sum_{m=1}^{M} p_m * n_{cm} \tag{4-1}$$

$$p_m = \sum_{i=1}^{S}\left(\frac{n_{cm}^{i}}{n_{cm}}\right)^2 \tag{4-2}$$

$$\overline{s} = \frac{1}{N}\sum_{i=1}^{S} s_i * n_{si} \tag{4-3}$$

$$s_i = \sum_{i=1}^{M}\left(\frac{n_{cm}^{i}}{n_{si}}\right)^2 \tag{4-4}$$

N 表示语音的总数量，S 表示语音中不同说话人的数目，M 表示聚类结束时的类数，n_{cm} 表示第 m 个类中语音的数量，n_{cm}^{i} 表示第 m 个类中由第 i 个说话人发出的语音数量，则 $\overline{p}$ 表示聚类的平均类纯度，p_m 表示第 m 类的类纯度，

n_{si}表示第 i 个说话人的语音数量，s_i表示第 i 个说话人的说话人纯度，$\bar{s}$表示平均说话人纯度。

类纯度也经常使用 RAND 评价参数[124,125]，RAND 表示的是不正确聚类的程度，使用不同类中来自同一说话人语音段的数量、或者同一个类中来自不同说话人语音段的数量来定义，公式为[124]：

$$R(M)=\frac{\sum_{m=1}^{M}(n_{cm})^2+\sum_{i=1}^{S}(n_{si})^2-2\sum_{m=1}^{M}\sum_{i=1}^{S}(n_{cm}^{i})^2}{\sum_{m=1}^{M}(n_{cm})^2+\sum_{i=1}^{S}(n_{si})^2}\times 100\% \qquad (4\text{-}5)$$

式（4-5）中，$R(M)$的值越小说明聚类的效果越好，理想的说话人聚类结果是 $R(M)=0$。

4.3　基于类纯度约束的说话人聚类算法

说话人聚类问题可分解成三个子问题，一是两段语音之间相似性的度量；二是语音合并准则；三是聚类停止条件。相似性度量的精度会直接影响聚类的性能，合并准则直接影响类纯度，停止条件则关系到聚类速度同时对性能也有一定的影响。本章主要针对第二个子问题，从降低说话人聚类结果对说话人辨认的影响出发，提出一种基于类纯度约束的说话人聚类算法（Class Purity Criterion based Speaekr Clustering, CPCSP），使用本文第 2 章中提出的基于 RSM 的距离度量准则来提高语音相似性度量的精度。该算法的基本思想是使聚类后类纯度高且类内语音总长度满足说话人辨认系统最短辨认长度要求的类尽可能多（最短辨认长度要求是指说话人辨认系统取得稳定性能对语音长度的要求），降低不同说话人的语音被聚到一类内的概率，下面分两部分介绍算法，首先分析聚类结果对说话人辨认的影响，然后详细介绍算法。

4.3.1　聚类结果的影响分析

从评价聚类结果的角度出发，聚类的理想结果是类纯度和说话人纯度都为 1，但在说话人检测当中由于分割点中误警及说话人平均发音长度较短，聚类语音段的长度较短，很难实现理想聚类结果。本文首先借助 RAND[124]参数来分析在说话人检测中聚类结果对最终说话人检测性能的影响程度，然后在不影响说话人检测性能的前提下确定重新说话人聚类的目标。

说话人聚类的理想结果是聚类数等于说话人数，每一类只包含一个说话人的语音，$M=S$ 且：

$$\begin{pmatrix} n_{c1}^{1} & n_{c2}^{1} & \cdots & n_{cM}^{1} \\ n_{c1}^{2} & n_{c2}^{2} & \cdots & n_{cM}^{2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{c1}^{S} & n_{c2}^{S} & \cdots & n_{cM}^{S} \end{pmatrix} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_S \end{pmatrix}_{S\times S} \tag{4-6}$$

其中，$n_k = n_{sk} = n_{ck}$。此时，式（4-5）经过推导可以得出 $R(M)=0$，为最小值[124]。

如果某两个说话人的语音被聚成了一类，即 $M=S-1$，假设是第 k、S 两个说话人被聚成了一类，则式（4-6）变为：

$$\begin{pmatrix} n_{c1}^{1} & n_{c2}^{1} & \cdots & n_{cM}^{1} \\ n_{c1}^{2} & n_{c2}^{2} & \cdots & n_{cM}^{2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{c1}^{S} & n_{c2}^{S} & \cdots & n_{cM}^{S} \end{pmatrix} = \begin{pmatrix} n_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & n_{S-1} \\ 0 & 0 & \cdots & n_S & \cdots & 0 \end{pmatrix}_{S\times(S-1)} \tag{4-7}$$

此时，式（4-5）经过推导可以得出 $R(M)=R(S-1)>0$，聚类效果变差，可见，如果 k、S 两个说话人中一个或者两个是目标说话人，就会发生目标说话人的语音被淹没在其他说话人语音中，很可能导致目标说话人漏检，从而对后面的说话人辨认造成严重影响。如果最终聚类的类数少于说话人的数目，对说话人检测的影响较大。

如果某一个说话人的语音被聚成了两类，即 $M=S+1$，假设是第 k 个类被分成了 k 和 S+1 两类，则式（4-6）变为：

$$\begin{pmatrix} n_{c1}^{1} & n_{c2}^{1} & \cdots & n_{cM}^{1} \\ n_{c1}^{2} & n_{c2}^{2} & \cdots & n_{cM}^{2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{c1}^{S} & n_{c2}^{S} & \cdots & n_{cM}^{S} \end{pmatrix} = \begin{pmatrix} n_1 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & n_2 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & n_{kk} & \cdots & 0 & n_{(S+1)k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & n_S & 0 \end{pmatrix}_{S\times(S+1)} \tag{4-8}$$

其中 $n_{kk}+n_{(S+1)k}=n_k$。此时，式（4-5）经过推导可以得出 $R(M)=R(S+1)>0$，聚类效果也在变差，但没有发生目标说话人的语音被其他说话人语音淹没，对说话

人检测来说，仅仅增加了说话人辨认的次数，同一个说话人被辨认了两次。

如果聚类结果中每一个说话人的语音都被分成两部分，一部分是仅包含该说话人的语音，另一部分称为一般结果。聚类结果可以表示为：

$$\left(\begin{array}{cccc|cccc} n_{c1}^{1} & n_{c2}^{1} & \cdots & n_{cM}^{1} & n_1 & 0 & \cdots & 0 \\ n_{c1}^{2} & n_{c2}^{2} & \cdots & n_{cM}^{2} & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{c1}^{S} & n_{c2}^{S} & \cdots & n_{cM}^{S} & 0 & 0 & \cdots & n_S \end{array}\right) \tag{4-9}$$

说话人检测中说话人聚类目的是增加语音长度，因此优先满足式（4-9）中分隔线右侧的 $n_1, n_2, \ldots, n_S$ 达到说话人辨认系统的最短辨认长度要求，对说话人检测的影响也会较小。

综合以上分析，可以对聚类的目标进行修改。将类内语音总长度达到最短辨认长度要求且类纯度高的类称为有效类，以有效类数量最大作为聚类目标。

4.3.2　算法描述

4.3.2.1　HAC 算法合并准则分析

在 HAC 算法中，首先寻找距离最近的两段语音，如果距离小于事先设定好的阈值，则将这两段语音合并，并未考虑语音合并对类纯度的影响，如图 4.2 所示：

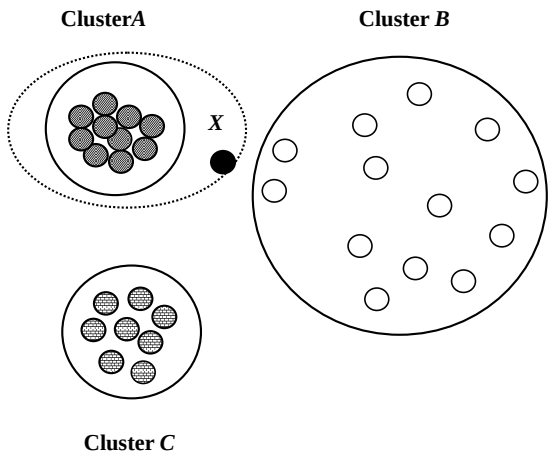


图 4.2 HAC 算法合并示意图

图 4.2 中，语音 X 与 A 类的距离最小，按照 HAC 算法，只要该距离小于设定好的阈值则语音 X 便会与 A 类合并，但合并会使 A 类的类内离散度急剧变大，而类内离散度与类纯度负相关，由此可见，如果阈值设置的不合理很可能导致不同

说话人的语音被聚在同一类，HAC 算法的聚类效果与阈值密切相关。更合理的结果是 X 不与 A 类合并，将 A 类单独归为一类，X 划分为一个新类，如图 4.3 所示。

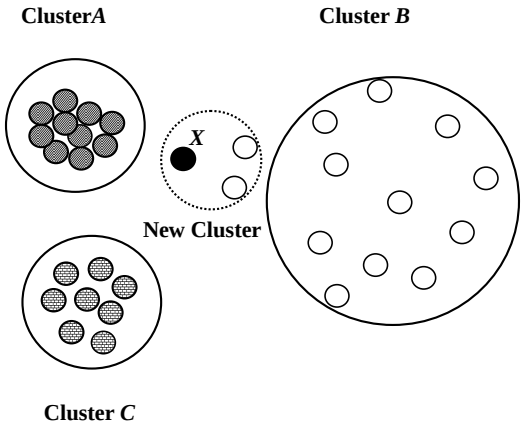


图 4.3 基于类纯度约束的说话人聚类示意图

4.3.2.2 基于类纯度约束的聚类算法描述

为了降低聚类结果对于说话人辨认的影响，必须提高聚类结果的类纯度，目前已有的改善类纯度方法大都是从提高平均类纯度出发[125-129]，但在说话人检测当中，平均类纯度的高低与说话人检测结果的好坏并没有直接的关系，可能平均类纯度很高但还是存在目标说话人语音被其他说话人语音淹没，导致说话人检测的效果变差。

由于在聚类过程当中无法得到某一类的类纯度值，而类纯度与类内离散度负相关，类内离散度越大类纯度越低。因此本文使用类内离散度来估计类纯度用来决定合并还是划分新类。定义类内离散度为：

$$w=\frac{1}{C_N}\sum_{i=1}^{C_N} d\left(X_i,\overline{X_i}\right) \qquad (4\text{-}10)$$

式（4-10）中，C_N 为某类的类内语音段的数量，X_i 是该类内的第 i 段语音，$\overline{X_i}$ 代表第 i 段语音以外的其他类内语音之和，$d\left(X_i,\overline{X_i}\right)$ 与第 2 章中式（2-31）相同，利用参考说话人模型度量语音 X_i 和 $\overline{X_i}$ 之间的距离。

如果一个语音段 Y 与 A 类的距离最小，但 A 类与语音段 Y 合并后使合并后的 A 类的类内离散度变大且超过某一阈值，则语音段 Y 单独分为一类；如果 A 类与语音段 Y 合并后 A 类的类内离散度变化小于阈值，则将语音段 Y 与 A 类合并。

根据 4.3.1 中的分析，在每一类的类纯度都较高的前提下，聚类数大于说话人数的聚类结果对说话人检测的影响较小。因此，在基于类纯度约束的说话人聚类算法中假设待聚类的语音中至少存在两个不同的说话人。首先从语音段中挑选最不可能属于同一说话人的两段语音，然后挑选与这两段语音距离最小且对类内离散度影响小的语音段与其合并，如果对类内离散度影响较大则将该段语音单独归为一类；如果某类语音达到了辨认语音长度要求，将该语音段单独归为一类不再参加后面的聚类过程，重复直至所有的语音都被处理。算法的详细步骤如下：

（1）定义两个类集合 C_I 和 C_S 并将其置为空，在所有的语音段中挑选语音长度达到最短辨认长度要求的语音，将这些语音段分别作为一类加入类集合 C_I；将所有语音段标记为“未归类”状态；最短辨认长度由说话人辨认系统的性能决定，本文实验中设为 5 秒。

（2）在“未归类”的语音段中长度超过 2 秒的语音中挑选相互间距离最大的两段语音，这两段语音分别作为一类加入类集合 C_S 中并将这两段语音标记为“已归类”，距离度量使用第 2 章中的基于 RSM 的度量准则，同时使用第 2 章中训练得到的 RSM 集合，记做 R_1；

（3）在 R_1 中分别挑选与类集合 C_S 中的语音距离最小的前 N 个参考说话人模型，将这些目标说话人模型作为一个新的 RSM 集合，记做 R_2；N 根据第 2 章的实验取 64；

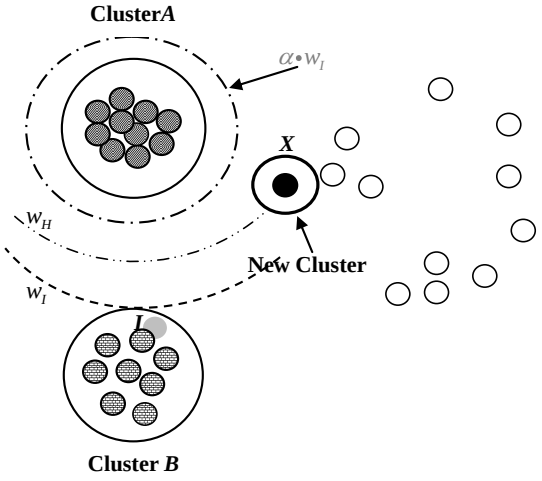


图 4.4 基于类内离散度约束的合并示意图

（4）利用 R_2 进行度量距离，分别计算“未归类”的语音段与类集合 C_S 中的语音段之间距离，找出距离最小的一段语音，将该语音段记为 H，将语音段 H 标记为“已归类”，C_{Si} 为 C_S 中与 H 距离最小的类，计算类 C_{Si} 加

入语音段 H 后的类内离散度 w_H；在类集合 C_S 中除类 C_{Si} 外的类的语音段中，挑选与类 C_{Si} 距离最小的语音段，该语音段记为 I，计算在类 C_{Si} 加入语音段 I 后类内离散度 w_I，如果满足 $w_H < \alpha * w_I$，则将 H 加入类 C_{Si} 中，否则 H 单独归为一类加入类集合 C_S 中；α 是一个可调节系数，$0<\alpha<1$。

（5）如果 C_S 中存在长度达到最短辨认长度要求的语音段，则将该语音段加入到集合 C_I 中并从 C_S 移除，在“未归类”的语音段中长度超过 2 秒的语音中与挑选与 C_S 距离最大的一段语音，将这段语音作为一类加入类集合 C_S 中，并将这段语音标记为“已归类”。

（6）如果所有的语音段均为“已归类”状态，则聚类结束；否则跳转到步骤 3。集合 C_I 即为聚类结果。

第 3 步中选取前 N 个距离最近的参考说话人模型的作用是为了增加第 4 步中距离度量的准确性，使用 R_2 度量距离是与第 2 章中的方法相同。第 4 步中，在算法开始运行时类内仅有一段语音段，无法进行类内离散度计算，本文通过将一段语音平均分成两段的办法来解决。

4.4　实验结果与分析

4.4.1　实验数据和设置

聚类实验是在 5 个不同的测试集下完成的，测试集 1、2、3 中参加聚类的每段语音都是单说话人语音，测试集 1、2 和 3 中的语音段平均长度分别为 2 秒、5 秒和 8 秒。测试集 1、2、3 是通过 NIST SRE 2006 中数据进行处理之后得到，挑选了男女各 100 位说话人的语音，每人一段 3 分钟长的语音，分割成长度不等的多段语音，长度从 0.5 秒到 10 秒不等。3 个测试集分别记做 Set1、Set2 和 Set3，每个测试集分成若干小集合，每个小集合包含的语音段数和说话人数都不一定相同。测试集 4、5 分别是第 2 章中实验五中 BNEWS 和 SWBD 数据库的分割结果。

聚类实验中比较了三个系统：基于 GLR 和 HAC 算法的系统；基于 RSM 和 HAC 算法的系统；基于 RSM 和 CPCSP 的系统；基于 GLR 的说话人分割系统作为基线系统。CPCSP 表示基于类纯度约束的聚类算法。

4.4.2　实验结果与分析

实验一　聚类性能比较

表 4.1 聚类性能比较

算法 / *pure*(%) / 测试集	GLR+HAC	RSM+HAC	RSM+CPCSP
Set1	78.5	79.6	81.2
Set2	82.5	84.5	86.3
Set3	84.9	87.4	89.5

表 4.1 中 *pure* 是指有效类的语音段长度和与总长度和之比。该实验验证了在不同测试集下，不同的说话人聚类算法获得有效类的能力，有效类的数量越多说明性能越好。GLR+HAC 和 RSM+HAC 算法使用阈值作为停止条件。由表 4.1 可以看出，本文提出的基于类纯度约束的说话人聚类算法 CPCSP 与 HAC 算法相比，得到的有效类语音的比例增大，在三个测试集合上，与基线系统相比分别提高了 2.7%、3.8%和 4.6%。

表 4.2 分割对聚类的影响

算法 / *pure*(%) / 测试集	GLR+HAC	RSM+HAC	RSM+CPCSP
BNEWS+DISTBIC	71.3	72.5	73.9
BNEWS+RSM	76.6	77.9	80.3
SWBD+DISTBIC	60.1	60.5	61.7
SWBD+RSM	60.8	61.0	62.8

表 4.2 中是 BNEWS+DISTBIC 表示使用 BNEWS 数据库的 DISTBIC 分割算法的分割结果作为聚类的输入语音。可以看出 CPCSP 算法的聚类性能要优于 HAC 算法，在四种聚类数据库下，有效类语音的比例均得到了提高。但由于分割中存在漏检导致聚类语音中含有一定程度的多说话人语音，与测试集 2 和测试集 3 相比，聚类性能变差。

实验二　聚类时间比较实验

在大规模目标说话人检测系统当中，算法速度也是一个相当重要的评价参数，因此本文对聚类算法的时间消耗进行了对比。

表 4.3 聚类消耗时间表

算法 / *F* / 测试集	GLR+HAC	RSM+HAC	RSM+CPCSP
Set1	1	0.39	0.35
Set2	1	0.28	0.21
Set3	1	0.30	0.23

表 4.3 中 *F* 是加速因子，是算法的聚类时间与 GLR+HAC 算法的聚类时间之比。在不同的测试集合上，使用基于 RSM 的度量准则比使用 GLR 要快，GLR 需要将语音训练成高斯混合模型，通常都是使用 EM 算法，EM 算法是要多次迭代到达收敛才停止，每次判断两段语音是否合并都需要训练模型，这是很大的时间消耗。BIC 和 KL 距离在本文中虽然没有进行实验比较，但其算法的基础都是需要进行模型训练，时间消耗同 GLR 属于同一量级。而基于 RSM 的距离度量准则的时间消耗小很多，因为在基于 RSM 的距离度量算法中每一帧语音仅需在所有参考说话人模型上计算一次并保存即可，两段语音合并仅需对每帧似然分进行简单运算就可以得到合并后新语音对 RSM 的似然分，从而得到新语音的距离矢量。CPCSP 算法由于只要一类语音的长度达到辨认长度要求即被单独归为一类不再用来聚类，这相当于在聚类的过程当中降低样本数量，因此其时间消耗和 HAC 算法相比会有明显的下降。

实验三　聚类性能对说话人检测的影响

聚类结果中有效类的数量越多对说话人辨认的影响越小，对此也进行了实验验证。首先将聚类实验中的 200 位说话人作为目标说话人，对聚类结果进行说话人辨认，说话人辨认系统基于 GMM-UBM，UBM 同第 2 章，GMM 的混合数为 1,024，辨认时选择似然分最高的前三个目标说话人作为候选。使用召回率和准确率两个参数来衡量说话人检测的性能。召回率是被检测出的真实目标说话人与多说话人语音中真实目标说话人的数目之比；准确率是被检测出的真实目标说话人与全部检出的目标说话人数目之比。

召回率定义：

$$R=\frac{\text{被检出的真实说话人数目}}{\text{真实说话人数目}}\times 100\% \qquad (4\text{-}11)$$

准确率定义：

$$P=\frac{\text{被检出的真实说话人数目}}{\text{全部的被检出说话人数目}}\times 100\% \qquad (4\text{-}12)$$

表 4.4 聚类对说话人检测性能的影响

算法 / *R*/*P*(%) / 测试集	GLR+HAC	RSM+HAC	RSM+CPCSP
Set1	68.1/21.3	71.0/23.8	75.7/28.3
Set2	76.2/22.2	78.6/26.9	82.4/27.8
Set3	82.5/27.4	83.9/28.3	87.6/27.6

表 4.4 中 *R*、*P* 分别表示召回率和准确率。通过表 4.1 和表 4.4 对比分析，有效类的数量与检测性能近似成正比。CPCSP 算法在聚类时权衡了距离和类内离散度变化，有效类的数量更多，保证了真实目标说话人能够被正确检出。由于说话人辨认系统给出的答案是 3 候选，保证了召回率较高的同时也使得准确率较低。RSM+CPCSP 算法的目标说话人召回率较 GLR+HAC 算法在 3 种测试数据集上分别提高了 7.6%、6.2%和 5.1%。

表 4.5 说话人检测性能−Set1

分割算法	聚类算法	FAR(%)	MDR(%)	*pure*(%)	*R*(%)	*P*(%)
BIC	GLR+HAC	28.8	27.8	60.5	58.8	20.5
GLR	GLR+HAC	29.5	26.4	60.9	61.0	22.1
RSM	RSM+CPCSP	24.3	24.6	61.4	65.7	28.9

表 4.6 说话人检测性能−Set2

分割算法	聚类算法	FAR(%)	MDR(%)	*pure*(%)	*R*(%)	*P*(%)
BIC	GLR+HAC	23.8	16.1	73.7	66.3	23.7
GLR	GLR+HAC	24.7	16.3	74.2	67.2	25.3
RSM	RSM+CPCSP	21.7	11.2	76.0	73.5	30.6

表 4.7 说话人检测性能−Set3

分割算法	聚类算法	FAR(%)	MDR(%)	*pure*(%)	*R*(%)	*P*(%)
BIC	GLR+HAC	21.4	14.9	80.7	75.2	26.3
GLR	GLR+HAC	23.6	15.1	80.5	76.5	27.5
RSM	RSM+CPCSP	20.9	10.3	81.6	85.6	29.2

表 4.5、4.6、4.7 中数据是对多说话人语音进行分割、聚类，然后进行辨认的实验结果。多说话人语音是对聚类实验数据进行一定的处理得到，将聚类用的语音段按照随机顺序进行拼接得到多说话人语音，测试集合 Set1 的窗宽取 1 秒，测试集合 Set2 和 Set3 的窗宽取 2 秒，窗移都是 0.3 秒。

由于分割中存在的漏检使得聚类语音中存在多说话人语音，导致聚类性能变差，从而使得说话人检测的结果与表 4.4 中的结果相比均有不同程度的下降，但 RSM 分割算法和 CPCSP 算法的组合在检测性能上还是较 BIC、GLR 和 HAC 的组合有明显改善；Set1 中包含较多的短发音，使得分割中存在的漏检较多，导致聚类和检测性能较差；Set2 和 Set3 中说话人平均一次发音长度较长，分割的难度在下降，分割效果逐渐变好，从而使得聚类结果和检测性能也随之变好。

4.5　小结

本章首先分析了在说话人检测当中说话人聚类对目标说话人辨认的影响，提出了基于类纯度约束的说话人聚类算法。该算法使用 RSM 度量语音间距离，以保持类纯度最大作为合并的约束条件。合并时不仅考虑语音段与类内已有语音的距

离，而且考虑合并对于类纯度的影响，以类内离散度评估类纯度。如果合并对于类内离散度的影响超过阈值，则停止合并而将该段语音单独分类；如果类内语音的总长度达到最短辨认长度要求，则该类单独归为一类不再继续聚类。在 NIST SRE 2006 数据库上，在语音段平均长度分别为 2 秒、5 秒和 8 秒的条件下，与传统的 HAC 算法比较，有效语音的比例分别提高了 2.7%、3.8%和 4.6%；目标说话人检测的召回率分别提高了 7.6%、6.2%和 5.1%。

第5章　基于双层结构的说话人快速辨认算法

在大规模目标说话人检测的处理流程中，将输入的多说话人语音进行分割聚类处理得到多段单说话人语音后，使用单说话人辨认技术检测是否有目标说话人的发音。大规模目标说话人对于说话人辨认的速度影响很大，目标说话人的数量与辨认时间成正比，而说话人检测系统一般都有严格的速度要求，因此必须进行加速处理，并且要在保持辨认准确率的前提下，这是本章的主要研究内容。

GMM-UBM[69]和 GMM-SVM[70]是两种最常用的说话人辨认方法，但传统的基于 GMM-UBM 或 GMM-SVM 的说话人辨认系统的辨认速度在目标说话人的规模很大情况下辨认速度很慢[4]。基于 GMM-SVM 的说话人辨认系统，需要将测试语音在 UBM 上进行自适应得到高斯混合模型作为 SVM 的输入，自适应过程的运算速度很慢且很难改进，本章研究如何提高基于 GMM-UBM 的说话人辨认系统的辨认速度。

本章的内容安排如下：5.1 简单介绍常用的说话人快速辨认算法，并给出相应的分析；5.2 介绍提出的基于参考说话人模型和双层结构的快速辨认算法；5.3 给出实验结果和分析；5.4 是本章的小结。

5.1　常用的说话人快速辨认算法

在第 1 章绪论中分析过，基于 GMM-UBM 的说话人辨认系统通常从以下三个方面开展研究，一是快速挑选核心分布；二是采用下采样方法压缩输入语音的特征向量数量，使用部分特征来代替全部特征；三是目标说话人剪枝，利用剪枝减少目标说话人的数量从而降低运算时间。

这些算法在一定条件下都能够提高了说话人辨认的速度而且性能下降很小，快速挑选核心分布算法单从挑选核心分布来讲，其加速性能非常好，但在大规模目标说话人条件下，挑选核心分布的计算量仅占说话人辨认总计算量的很小的一部分，因而其对整个辨认来说加速贡献很小；下采样的方法在大规模目标说话人检测系统中并不是十分适用，但其采样语音的似然分与全部语音的似然分之间存在一定的波动，算法的稳定性存在欠缺，而且由于进行辨认的特征往往是去除了静音的有效语音，相邻帧之间的相关性并不能保证较大；目标说话人剪枝算法对提高辨认速度的作用较大，下面简要介绍两种常用的基于目标说话人聚类的剪枝算法。

5.1.1　HSI算法

Sun 等 2005[96]提出的基于 GMM 的说话人辨认系统中使用目标说话人模型聚类将目标说话人集合组织成两层的层次结构，利用层次结构和剪枝技术进行说话人快速辨认。分为目标说话人聚类、重估类模型和剪枝辨认三个步骤。

目标说话人聚类。使用了迭代自组织数据分析算法（Iterative Self-Organizing Data Analysis Technique Algorithm, ISODATA）[97]进行说话人模型聚类。假设有两个目标说话人，λ_1和λ_2分别为其对应的 GMM，混合数分别为 H 和 L，二者之间距离为：

$$\varepsilon = \left(\mu_1^i - \mu_2^j\right)^2 \tag{5-1}$$

$$d_{ij} = \frac{\sigma_1^j}{\sigma_2^j} + \frac{\sigma_2^j}{\sigma_1^j} + \frac{\varepsilon}{\sigma_2^j} + \frac{\varepsilon}{\sigma_1^j} \tag{5-2}$$

$$d\left(\lambda_1, \lambda_2\right) = \sum_{i=1}^{H} w_i^1 \min_j d_{ij} + \sum_{j=1}^{L} w_j^2 \min_i d_{ij} \tag{5-3}$$

其中，d_{ij}是模型λ_1中第 i 个混合与模型λ_2中第 j 个混合之间的距离，$d\left(\lambda_1, \lambda_2\right)$是模型$\lambda_1$与模型$\lambda_2$之间的距离，$w_i^1$是模型$\lambda_1$中第 i 个混合的权重，w_j^2是模型λ_2中第 j 个混合的权重。

重估类模型。ISODATA 聚类结束后，将属于同一类的目标说话人的数据混合在一起重新训练一个新的 GMM 模型，使用该模型作为类模型。结构分成顶层和底层两部分，顶层中的节点为各类的类模型，底层中的节点为所有目标说话人的模型，每个底层节点在顶层中的父节点为该节点对应目标说话人所属类的顶层节点。

剪枝辨认。输入语音首先对顶层中各个节点即各个类模型计算似然分，剪枝掉似然分较低的节点仅保留似然分最高的节点，输入语音对该节点的底层子节点对应的目标说话人模型计算似然分，根据似然分的高低进行说话人辨认。

在包含 40 个目标说话人的测试集合上的实验表明，HSI 算法仅用了未加速的基线系统 30.3%的运算时间，而辨认性能仅稍有下降。

5.1.2　SMC算法

Apsingekar 和 Leon 2007 提出[4]，该算法的基本思想与层次结构的说话人辨认算法类似，主要过程分为目标说话人聚类和剪枝辨认两个步骤。与 HSI 算法不同的是，在聚类之后不进行类模型重估而是直接使用聚类中心作为类模型。

目标说话人聚类。使用了 K-Means 聚类算法[98,99]进行聚类，并分别使用欧式距离、KL 距离和对数似然分等多种距离度量方式进行了聚类实验，其中对数似然分的性能最好。假设 λ_1 为一个目标说话人的 GMM，$X=\{x_1,x_2,...,x_T\}$ 为训练 λ_1 的语音的特征，λ_c^i 为第 i 类中心对应的 GMM，二者之间的对数似然分的计算公式如下：

$$\begin{aligned} d\left(\lambda_1,\lambda_c^i\right) &= -\frac{1}{T}\sum_{m=1}^{T}\log\left(p\left(x_m\mid\lambda_c^i\right)\right) \\ &= \frac{D}{2}\log\left(2\pi\right)+\frac{1}{2}\log\left|\Sigma\right| \\ &\quad +\frac{1}{2T}\sum_{m=1}^{T}\left[\left(x_m-\mu\right)^t\Sigma^{-1}\left(x_m-\mu\right)\right] \end{aligned} \tag{5-4}$$

剪枝辨认。输入语音首先对各类的类中心模型计算对数似然分并排序，剪枝去掉类得分较低的类，选择类模型得分最高的类，使用输入语音对属于该类的目标说话人的模型计算似然分，根据似然分的高低进行说话人辨认。

在 TIMIT[113]，NTIMIT[129]和 NIST SRE 2002[5]三个数据库中进行了实验，三个数据库中目标说话人的个数分别为 630、630 和 330。在 TIMIT，NTIMIT 数据库的实验中目标说话人的聚类数是 100；在 NIST SRE 2002 数据库的实验中聚类数是 50。基于 GMM-UBM 的基线系统在三个数据库上的辨认准确率分别为 99.68%，69.37%和 89.39%。辨认时分别挑选了类中心的对数似然分在前 10%和前 20%的类分别进行了辨认实验。挑选前 10%的类时，算法的加速因子大约为 8.7，对三个数据库的辨认性能分别下降了 0.95%，2.2%和 1.4%；挑选前 20%的类时，算法的加速因子大约为 4.4，对三个数据库的辨认性能均没有下降。

5.1.3　现有算法分析

目标说话人剪枝算法 HSI 和 SMC 都取得了较好的加速效果，而且辨认性能的下降很小甚至没有。但是两种算法在实验中使用的目标说话人的数量都不大，HSI 使用了 40 个目标说话人，SMC 最大的数据库有 630 个目标说话人，在更大规模的目标说话人条件下其性能是否会产生变化。一般来说，随着目标说话人数量的增多，目标说话人的声学特性之间更容易发生交叠，目标说话人聚类的效果通常会

变差，即类间区分性往往会降低。类间区分性的降低必然有更多的目标说话人位于类间边界的位置，即更多的目标说话人数量容易发生类选择上的混淆。当待辨认语音位于容易发生混淆的类间边界时，很可能会由于类的选择错误而致使真正的目标说话人无法被正确检测到。SMC 通过挑选前 20%的类进行辨认来减弱这一影响保持性能，但增加辨认的类数会导致运算性能的提高幅度下降，SMC 的加速因子从 10%的 8.7 降低到 20%的 4.4。

SMC 和 HSI 算法性能受目标说话人规模的影响，其根本原因在于剪枝后保留的目标说话人与聚类中心相似程度较高，而未必与待辨认语音相似程度较高，当待辨认语音与处于易混淆的两类或多类间的边界时，SMC 和 HSI 仅保留了某一类中的目标说话人，而理想的结果是保留这两类边界处与待辨认语音相似程度高的目标说话人，即保留与待辨认语音相似程度高的目标说话人。

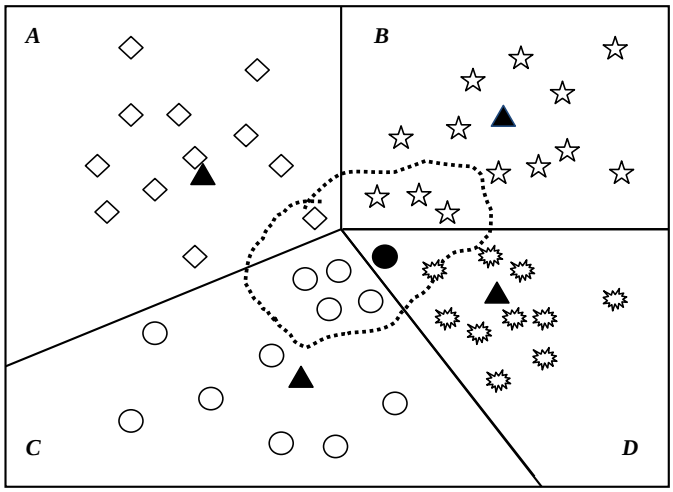


图 5.1 HSI 和 SMC 算法的目标说话人剪枝示意图

例如图 5.1 中，*A*、*B*、*C* 和 *D* 四类是目标说话人聚类结果中的一部分，黑色实心三角形代表聚类中心，黑色实心圆形表示待辨认语音，待辨认语音与四个类中心的距离大小的顺序为 $D<B<A<C$，按照 SMC 或 HSI 算法，如果挑选一类会选择 *D* 类中的目标说话人进行辨认，如果选择两类会选择 *D* 和 *B* 两类，而真实目标说话人也有较大可能存在于 *A* 类或 *C* 类中，选择类数多可以提高真实目标说话人的被检出的概率，但会使加速效果变差。而理想情况应该挑选被虚线包围的目标说话人进行辨认，这些目标说话人与待辨认语音更接近。

基于以上分析，本文提出了一种基于 RSM 和双层结构的说话人快速辨认算法，使用双层结构挑选与待辨认语音相似程度高的 RSM，借助这些 RSM 来度量待辨认语音与目标说话人的相似程度，剪枝去掉相似程度低的目标说话人，保留与待辨认语音相似程度高的目标说话人进行说话人辨认，从而提高辨认速度。

5.2　基于双层结构的说话人快速辨认算法

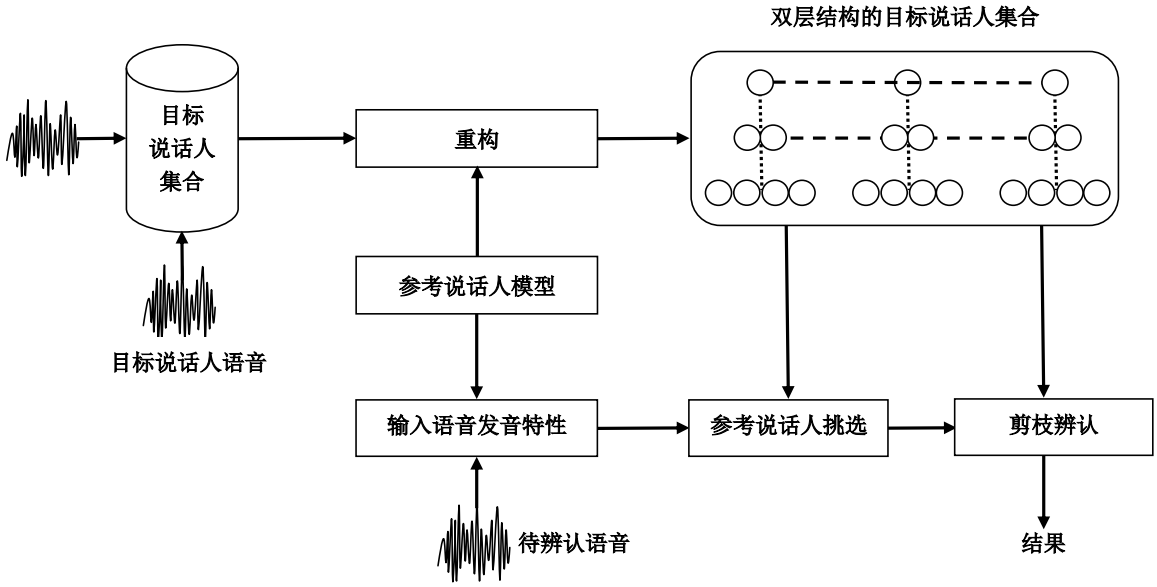


图 5.2 加速算法的流程示意图

从本文第 2 章中的结论可知，参考说话人模型可以用来度量两段语音之间的距离，那么同样可以度量待辨认语音与目标说话人语音之间的距离来评估二者的相似程度，目标说话人与参考说话人模型的距离可在训练阶段预先全部计算好，辨认时仅计算待辨认语音与参考说话人模型的距离、以及待辨认语音距离向量和目标说话人距离向量之间的距离，并将距离作为相似程度度量来选择最相似的一部分目标说话人进行说话人辨认。基于参考说话人模型的方法，待辨认语音需要在参考说话人模型上计算似然分，还要计算距离向量之间的距离，HSI 和 SMC 算法仅需使用待辨认语音对聚类中心计算似然分，相似程度的计算可看作一个额外的计算消耗，但这一计算的速度较快而且由于相似程度准确性高会使选择到的目标说话人能更准确的召回真正目标说话人。相比于 HSI 和 SMC 方法的选择多类，本算法能够在选择较少目标说话人数量的同时保证较高的真实目标说话人召回率。下面本文首先介绍基于参考说话人模型的剪枝算法并分析其加速性能，然后介绍双层结构并分析加速性能。

5.2.1　基于RSM的目标说话人剪枝算法

图 5.3 中，黑色实心三角形符号代表参考说话人模型，黑色空心圆形代表目标说话人模型，实心直线代表目标说话人的语音与参考说话人模型的距离；黑色实心圆形代表待辨认语音，黑色虚线代表待辨认语音对参考说话人模型的距离；根据待辨认语音与目标说话人在参考说话人模型上的距离的差异来挑选与待辨认语音最相似的那些目标说话人。

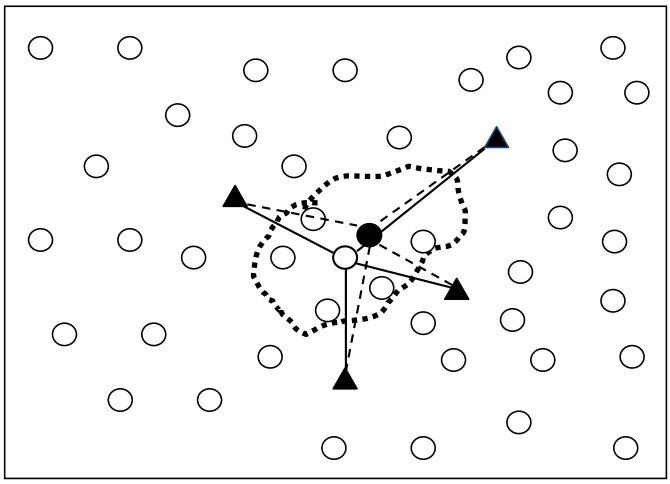


图 5.3 基于参考说话人模型的目标说话人剪枝算法示意图

5.2.1.1　算法描述

算法步骤的详细步骤如下：

训练阶段：

（1）RSM 训练方法与第 2 章中的方法相同。使用最小最大方法从目标说话人模型集合挑选 K 个模型作为初始中心，选择方法参见第 2 章。

（2）使用 K-means[98,99]算法将训练参考说话人模型的说话人模型聚成 K 类，得到 K 个聚类中心，这 K 个聚类中心即为 RSM。

（3）使用 K 个参考说话人模型对目标说话人进行矢量量化，使用训练目标说话人模型的语音对 K 个参考说话人模型计算似然分，每个目标说话人得到 K 个似然分。

（4）将每个目标说话人的 K 个似然分拼接成一个 K 维偏差向量 V_T，计算 V_T 的均值和标准差并保存。

辨认阶段：

（5）使用 K 个参考说话人模型对待辨认语音进行量化，计算待辨认语音对 K 个参考说话人的似然分，得到 K 维的偏差向量 V_I。

（6）计算 V_I 与所有 V_T 的相关系数（利用 4 中预先计算好的均值和标准差可以节约计算时间），挑选相关系数最大的前 L 个目标说话人模型。

（7）计算待辨认语音在步骤 6 中得到的 L 个目标说话人模型上的似然分，使用 ZNorm 算法[130]进行的分数规整，根据得分高低辨认说话人。

5.2.1.2　加速性能分析

假设 UBM 的混合数为 M，核心分布的数量为 N，目标说话人的数量为 H，待辨认语音的数量为 B，每个待辨认语音的特征数量为 T，未加速时系统需要计算似然分的单高斯分布的数量为 M_N，M_N 计算公式如下：

$$M_N = \sum_{i=1}^{B} T_i \left(M + NH \right) \tag{5-5}$$

其中 T_i 是第 i 个待辨认语音的特征数量。

设参考说话人模型的数量为 K，剪枝后保留的目标说话人数量为 L，加速后系统需要计算似然分的单高斯分布的数量为 M_S，M_S 计算公式如下：

$$M_S = \sum_{i=1}^{B} T_i \left(M + N\left(K + L \right) \right) + BT_C \tag{5-6}$$

其中，T_C 为 H 个 K*1 维向量之间相关系数的计算时间除以单高斯分布似然分计算时间的值，即将 H 个相关系数的计算时间折算成单高斯分布的计算数量，加速因子为：

$$F_S = \frac{M_N}{M_S} = \frac{\sum_{i=1}^{B} T_i \left(M + NH \right)}{\sum_{i=1}^{B} T_i \left(M + N\left(K + L \right) \right) + BT_C} \tag{5-7}$$

为便于分析，设所有待辨认语音的特征数量均相同为 T，则式（5-7）可以化简为：

$$F_S = \frac{M_N}{M_S} = \frac{M + NH}{M + N\left(K + L \right) + \dfrac{T_C}{T}} \tag{5-8}$$

T_C 的值受目标说话人的数量影响，但其计算仅包含基本的加法和乘法运算，与似然分运算中的对数运算和指数运算相比，耗时很小，且 T 的值越大，T_C 对加速因子的影响越小。由于 T_C 的存在，使得在参考说话人数量等于 SMC 或 HSI 算法的聚类数时，以及保留相同的目标说话人条件下，本算法加速因子较 SMC 或 HSI 算法会略有下降，但保留的目标说话人与待辨认语音的相似程度更高，因而本算法的辨认性能会比 SMC 或 HSI 算法要好。

设某快速挑选核心分布算法挑选到核心分布时每帧语音需计算 Q 个单高斯分布的似然分，使用该加速算法的系统需要计算似然分的单高斯分布的数量为 M_Q，M_Q 的计算公式如下：

$$M_Q = \sum_{i=1}^{B} T_i \left(Q + NH \right) \tag{5-9}$$

快速挑选核心分布算法的加速因子：

$$F_Q = \frac{M_N}{M_Q} = \frac{M + NH}{Q + NH} \tag{5-10}$$

在 M=1,024，N=4，H=5,000，K=256，L=300 的条件下，基于参考说话人模型的剪枝算法（Reference Speaker Model based Speaker Pruning, RSMSP）算法的加速因子 F_S 约为 6.5，目标说话人的数量越多，K 与 L 的和越小，算法的加速因子越大，但加速因子过大会导致性能下降过快，一般在保证辨认性能的前提下使加速因子最大。在 Q=64 时，快速挑选核心分布算法的加速因子 F_Q 仅为 1.05，由此可见，在大规模目标说话人条件下，仅使用快速挑选核心分布算法对辨认的加速作用很有限，只有当目标说话人数量 H 较小时，快速挑选核心算法对整个辨认的作用才会比较显著。如果将快速挑选核心分布算法与目标说话人剪枝算法融合则加速因子：

$$F_{SQ} = \frac{M_N}{M_Q} = \frac{M + NH}{Q + N\left(K + L \right) + \dfrac{T_C}{T}} \tag{5-11}$$

在上述条件下，F_{SQ} 约为 9.2，快速挑选核心分布算法与目标说话人剪枝融合之后的加速效果比较好。

5.2.2　基于双层结构的目标说话人剪枝算法

由于多参考说话人模型覆盖的声学空间较广，而一般来说待辨认语音仅与一小部分参考说话人模型的距离较小而与大部分参考说话人模型的距离较大，这些较大的距离对于待辨认语音和目标说话人的区分能力较小，计算相似程度还会造成一定的精度影响同时增加计算时间消耗，因此考虑将与待辨认语音距离较大的参考说话人模型去除，使用距离较小的一部分参考说话人模型度量待辨认语音与目标说话人的相似程度，可以在一定程度上提高计算速度。基于此本文提出基于 RSM 和双层结构的说话人快速辨认算法，上层用来对 RSM 进行快速挑选并对目标说话人进行粗剪枝，下层用来对目标说话人进行精确剪枝和辨认。

5.2.2.1　算法描述

训练阶段：

（1）首先使用上一小节中的方法训练 K_D 个下层参考说话人模型，记做 DRSM；使用这 K_D 个参考说话人模型作为输入训练出 K_U 个上层参考说话人模型，记做 URSM，且 $K_D > K_U$。

（2）计算目标说话人与 DRSM 的偏差向量 V_T^D；计算目标说话人与 URSM 的偏差向量 V_T^U；计算 DRSM 与 URSM 的偏差向量 V_D^U。

辨认阶段：

（3）计算待辨认语音与 K_U 个 URSM 的似然分，并拼接成 K_U 维的上层偏差向量 V_I^U。

（4）计算 V_I^U 与所有 V_D^U 的相关性并按照大小进行降序排列，挑选出前 J 个 DRSM 并记录索引，并根据这 J 个 DRSM 的索引从 V_T^D 抽取对应的子偏差向量 V_T^{DJ}；计算 V_I^U 与所有 V_T^U 的相关性并按照大小进行降序排列，挑选出前 R 个目标说话人。

（5）计算待辨认语音与步骤 4 中挑选的 J 个下层参考说话人模型的偏差得到 J 维的下层偏差向量 V_I^{DJ}。

（6）从计算 V_I^{DJ} 与步骤 4 中挑选的 R 个目标说话人的 V_T^{DJ} 的相关性，并按照大小进行降序排列，挑选出相关性最大的前 L 个目标说话人。

（7）计算待辨认语音在步骤 6 中得到的 L 个目标说话人模型上的似然分，根据得分高低辨认说话人。

5.2.2.2　加速性能分析

设下层参考说话人模型数量为 K_D，上层参考说话人模型数量为 K_U，剪枝后保留的下层参考说话人数量为 J，目标说话人数量为 L，加速后系统需要计算似然分的单高斯分布的数量为 M_T，M_T 的计算公式为：

$$M_T=\sum_{i=1}^{B}T_i\left(M+N\left(K_U+J+L\right)\right)+B\left(T_C^{UH}+T_C^{UK_D}+T_C^{DR}\right) \qquad (5\text{-}12)$$

其中，T_C^{UH}、$T_C^{UK_D}$ 分别为 H 个、K_D 个上层偏差向量之间的相关系数的计算时间折算的单高斯分布似然分计算数量，T_C^{DR} 为 R 个下层偏差向量之间的相关系数的计算时间折算的单高斯分布似然分计算数量，则加速因子为：

$$F=\frac{M_N}{M_T}=\frac{\sum_{i=1}^{B}T_i\left(M+NH\right)}{\sum_{i=1}^{B}T_i\left(M+N\left(K_U+J+L\right)\right)+B\left(T_C^{UH}+T_C^{UK_D}+T_C^{DR}\right)} \qquad (5\text{-}13)$$

为便于分析，设所有待辨认语音的特征数量均相同为 T，则式（5-13）可化简为：

$$F=\frac{M_N}{M_T}=\frac{M+NH}{M+N\left(K_U+J+L\right)+\frac{\left(T_C^{UH}+T_C^{UK_D}+T_C^{DR}\right)}{T}} \qquad (5\text{-}14)$$

在 M=1,024，N=4，H=5,000，K_D=256，K_U=32，J=64，R=1000，L=300 的条件下，加速因子约为 8.1。融合快速挑选核心分布算法，取 Q=64，则加速因子约为 11.9。

$$F=\frac{M_N}{M_T}=\frac{M+NH}{Q+N\left(K_U+J+L\right)+\frac{\left(T_C^{UH}+T_C^{UK_D}+T_C^{DR}\right)}{T}} \qquad (5\text{-}15)$$

5.3　实验结果与分析

5.3.1　实验数据和设置

说话人辨认加速实验的数据来自 CCC 的 VPR-2C2005-6000[100]，语音是在电话信道下录制，采样频率 8KHz，采样精度 8 位，单声道录音，选择 5,200

个说话人作为目标说话人，目标说话人的训练语音为 30 秒，辨认语音由 5,200 个目标说话人和 1,000 集外说话人的语音组成，每个待辨认说话人有 1 条语音，平均待辨认语音长度为 5 秒。实验中的基线系统基于 GMM-UBM，UBM 和说话人模型的混合数均为 1,024。

评价一个说话人辨认系统性能的主要参数是辨认正确率，辨认正确率有首选正确率（*Top*-1）和前 *N* 选正确率（*Top-N*），本文中辨认正确率选择 *Top-N*，*N* 取 3，参见式（5-16）。对于本文研究的开集的说话人辨认，还要判断出待辨认语音是否为集外说话人，实验中是根据预先设定好的辨认阈值来判断待辨认语音是目标说话人还是集外说话人，如果待辨认语音与目标说话人的相似程度超过阈值则判定为目标说话人，否则为集外说话人。如果待辨认语音判定为目标说话人，则挑选相似程度最高且超过阈值的 3 位目标说话人作为答案，如果不足 3 位则取所有相似程度超过阈值的目标说话人作为答案。辨认阈值通过开发集获得，实验中取 DET 曲线上 EER 点所对应的阈值来作为辨认阈值。

$$Top\text{-}3=\frac{R}{A}*100\% \qquad (5\text{-}16)$$

式（5-16）中，*A* 是总辨认次数，*R* 是辨认正确的次数。如果待辨认语音为集外说话人，“集外说话人”为辨认正确的答案；如果待辨认语音为集内说话人，真实说话人出现在候选目标说话人当中则为辨认正确；其他情况均为辨认错误。

5.3.2　实验结果与分析

实验一　目标说话人数量对聚类效果的影响

实验中聚类用的说话人是从实验数据库 6,200 个说话人中随机抽取而得到的子集，聚类数目是根据目标说话人数量选择多个聚类数然后从聚类结果中选取最优的类数，评价聚类效果的参数是类间平均距离与类内平均距离的比值，比值越大则聚类效果越好反之则越差。

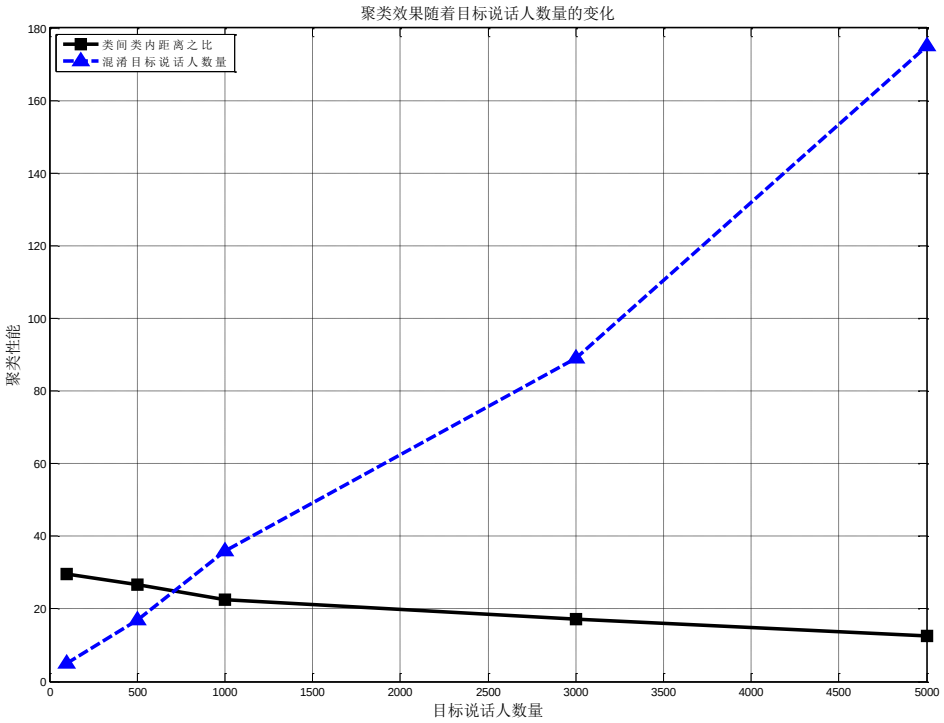


图 5.4 聚类效果与目标说话人数量的关系

图 5.4 中是多次实验的平均值。混淆目标说话人数量是聚类之后类间容易混淆的目标说话人，计算目标说话人与 K 个类中心的距离并排序，如果最小的 3 个距离值之间的差异较小则该目标说话人定义为混淆目标说话人。从图 5.4 中可以看到，随着目标说话人数量的增多，类间区分性在降低，混淆目标说话人的数量在增加。这就说明随着目标说话人数量的增多类间的混淆程度在增加，聚类效果在变差，对 SMC 和 HSI 方法的影响也会随之加大。

实验二　DRSM 数量的选择

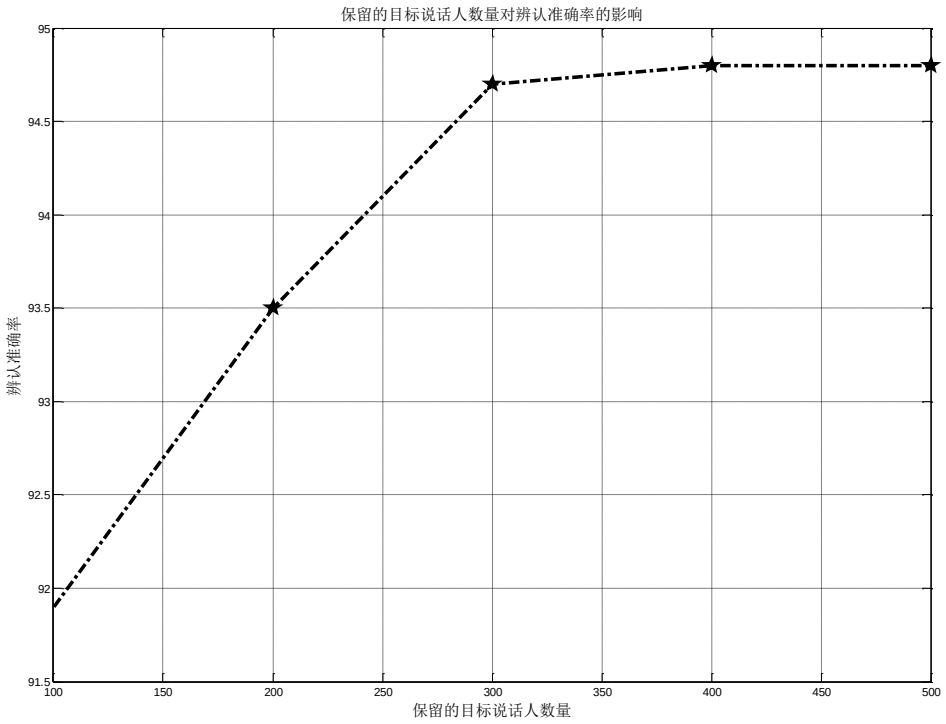


图 5.5 保留目标说话人数量对辨认准确率的影响

剪枝掉的目标说话人数量越多，保留的目标说话人越少，系统加速性能越好，但由于相似性度量精度的问题，剪枝过多会造成真正目标说话人被剪枝从而导致辨认准确率下降，出于加速因子和辨认准确率两方面平衡的考虑，在后续实验中参数 L 选择 300。加速因子是基线系统的运行时间与实验系统的运行时间之比。

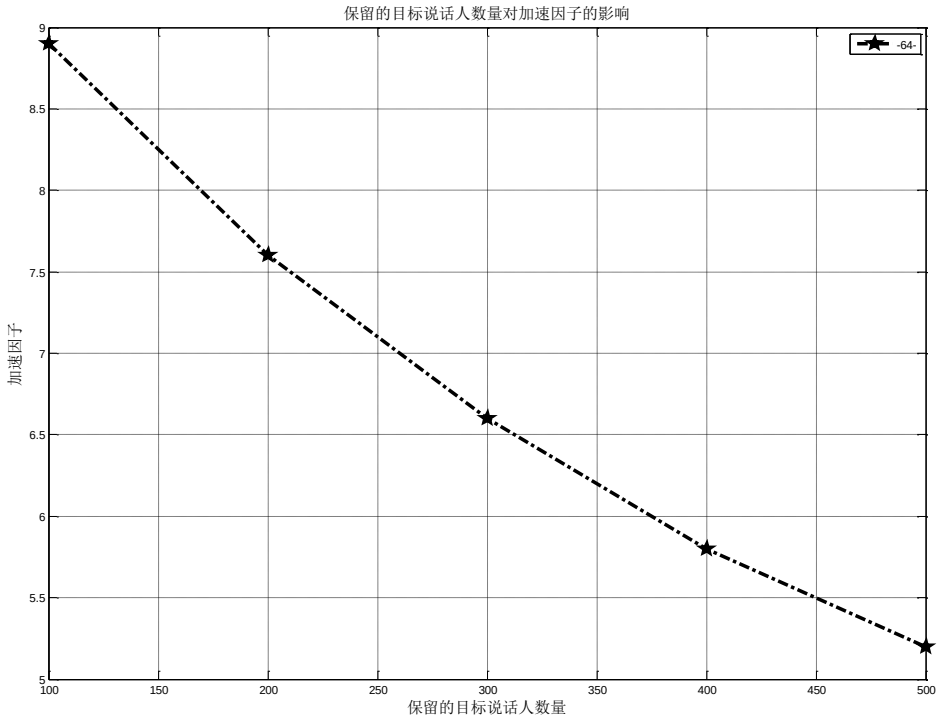


图 5.6 保留目标说话人数量对加速因子的影响

实验三　剪枝性能比较实验

表 5.1 剪枝性能比较

方法	*L*	*F*	*AE*	*Top*−3(%)
SMC	100	8.9	0.40	88.4
	200	7.6	0.58	89.3
	300	6.5	0.70	92.9
	400	5.8	0.81	93.1
	500	5.2	0.87	93.2
RSMSP	100	8.5	0.35	91.9
	200	7.3	0.47	93.5
	300	6.3	0.61	94.7
	400	5.6	0.69	94.8
	500	5.1	0.78	94.8

表 5.1 中，*AE* 表示平均偏差，描述保留的目标说话人与真实的目标说话人的平均相似程度。

$$AE = \frac{1}{L}\sum_{i=1}^{L} KL\left(\lambda_R^i, \lambda_T\right) \tag{5-17}$$

其中，λ_R^i 是第 i 个保留目标说话人的高斯混合模型，λ_T 是真实目标说话人的高斯混合模型，λ_R^i 与 λ_T 之间的 KL 在距离描述了保留目标说话人 i 与真实目标说话人的相似程度，KL 距离越小说明二者的相似程度越高，*AE* 小则说明保留的目标说话人当中与真实的目标说话人相似程度高的多，真实说话人被保留的概率越大，辨认性就越好。

在表 5.1 中，RSMSP 表示基于参考说话人模型的目标说话人剪枝算法，保留相同数量的目标说话人时 RSMSP 算法由于增加了相似性计算导致其运算时间较 SMC 算法增加，但其选择得到目标说话人集合与真实目标说话人的平均偏差更低。随着保留目标说话人数量的增多，相似性计算所用时间与整个辨认的总时间之比变小，二者加速因子的差异也在变小。SMC 算法的聚类数与 RSMSP 的参考说话人数量相同。

保留的目标说话人数量相同时 RSMSP 算法较 SMC 方法的运算时间有所增加，但其选择得到目标说话人集合与真实目标说话人的 *AE* 比 SMC 方法的 *AE* 小，RSMSP 方法保留较少的目标说话人可以达到 SMC 方法保留较多的目标说话人时的辨认准确率。

实验四　双层结构设计实验

URSM 用来挑选与待辨认语音相似程度高的 DRSM，以及用来对目标说话人模型进行粗剪枝，挑选与待辨认语音相似程度高的 DRSM 与本文第 2 章中的方法类似，通过实验来选择合适的值。表 5.2 中 K_U 是 URSM 的数量，J 是挑选的与待辨认语音相似程度最高的 DRSM 的数量，在加速因子和辨认准确率两方面进行一定的平衡之后选择 $K_U = 32$，$J = 64$。

表 5.2 双层结构设计实验

K_U	J	F	$Top-3$(%)
	32	8.8	93.3
16	64	8.5	93.5
	128	7.6	93.8
	32	8.6	94.1
32	64	8.2	94.5
	128	7.4	94.6
	32	8.2	94.2
64	64	7.8	94.6
	128	7.1	94.6

实验五　加速性能比较

表 5.3 算法性能比较

算法	F	$Top-3$(%)
Baseline	1.0	94.9
SMC	5.8	93.1
RSMSP	6.3	94.7
RSMSP+TL	8.2	94.5
SMC+TBKS	8.1	93.0
RSMSP+TBKS	9.3	94.5
RSMSP+TL+TBKS	11.9	94.2

RSMSP 剪枝算法与 SMC 算法相比，在运算速度的提高相同的条件下，可以取得更好的辨认效果，在使用双层结构剪枝算法后，其运算速度会得到进一步提高，RSMSP+TL 算法与 SMC 算法相比，运算时间降低 29.3%而辨认准确率提升 1.4%。RSMSP 算法与快速挑选核心分布算法 TBKS 的融合能够进一步提升辨认的运算速度，RSMSP+TL+TBKS 算法与 SMC+TBKS 算法相比，运算时间降低了 31.9%而辨认准确率提升 1.2%。

实验六　大规模目标说话人检测系统性能实验

实验中多说话人语音使用实验五中的辨认数据进行随机拼接得到，说话人平均一次发音长度为 5 秒，首先对其进行说话人分割聚类，然后进行说话人辨认，基线系统的分割算法使用 DISTBIC，聚类算法使用 HAC，辨认算法使用 SMC+TBKS。

表 5.4 算法性能比较

分割算法	聚类算法	辨认算法	*F*	*R*(%)	*P*(%)
DISTBIC	HAC	GMM-UBM	1	70.1	25.8
		SMC	5.8	69.0	23.4
		SMC+TBKS	8.1	68.1	22.6
RSM	CPCSP	GMM-UBM	1	78.3	29.1
		RSMSP	6.3	76.8	26.2
		RSMSP +TL	8.2	76.7	24.0
		RSMSP+TL+TBKS	11.9	74.9	23.5

与基线系统相比，本文算法的目标说话人召回率提高了 6.8%，运算时间降低了 31.9%。

5.4　小结

本章首先对现有基于 GMM-UBM 的说话人辨认快速算法进行了分析，针对其存在问题，提出一种基于参考说话人和双层结构的目标说话人剪枝算法，将目标说话人模型和参考说话人模型组织成双层结构，首先挑选与待辨认语音相似程度高的参考说话人模型，利用这些参考说话人快速度量目标说话人与待辨认语音的相似程度，并利用相似程度进行目标说话人剪枝。在 CCC VPR-2C2005-6000 数据库上，在 5,200 个目标说话人和 1,000 个集外说话人的开集辨认条件下，在基于 GMM-UBM 架构的说话人辨认系统中，

RSMSP 算法与 SMC 算法相比，运算时间降低 29.3%而辨认准确率提升 1.4%。RSMSP 算法与快速挑选核心分布算法 TBKS 的融合能够进一步提升辨认的运算速度，与 SMC 和 TBKS 的融合算法相比，运算时间降低了 31.9%而辨认准确率提升 1.2%。RSMSP 算法与 GMM-UBM 系统相比，在提高 8.2 倍运算速度的同时而辨认准确率仅相对下降 0.2%，与 TBKS 融合后可在提高 11.9 倍运算速度的同时而辨认准确率仅相对下降 0.7%；

使用本文提出的基于参考说话人的分割算法、基于类约束的聚类算法和基于参考说话人和双层结构的目标说话人剪枝算法的大规模目标说话人检测系统，相对于分割算法使用 DISTBIC 算法、聚类算法使用 HAC 算法，辨认算法使用 SMC+TBKS 的基线系统，目标说话人的召回率提高了 6.8%，运算时间降低 31.9%。

第6章　结论与展望

6.1　论文工作总结

随着科学技术的不断发展，生物特征识别技术在社会生活的各个领域的应用也越来越广泛和深入，在各种生物特征身份认证技术中，说话人识别技术由于其某些自身独特性质而具有不可替代性。如说话人识别技术对生物特征的采集不涉及敏感的个人隐私，用户接受性强；说话人识别所需的设备成本低廉甚至无需额外增加成本（如电话应用中），易于推广；在某些应用场景（如电话网络）中，说话人的语音特征是当前较少甚至是唯一可以轻易获取的生物特征；说话人语音中所蕴涵的丰富的韵律信息可以反应说话人意图的真实性[131]。说话人识别的研究始于 20 世纪 30 年代，并在近三十多年里取得了长足的进步，本文针对其中的一个研究任务，即大规模目标说话人检测中的若干难点和问题，在说话人分割聚类和说话人快速辨认方面进行了初步的探索和研究，提出了一些新的方法，并通过实验证明了其有效性，同时也为进一步深入研究打下了一定的基础。

概括来说，本文的工作重点和贡献主要体现在如下几个方面：

1．针对说话人分割中存在的距离度量精度问题，提出了一种**基于参考说话人模型的说话人分割算法**。该算法不需要将窗内语音训练成模型，避免了数据较少、模型不精确对距离度量带来的影响，它利用参考说话人模型来度量两段语音之间的相似程度。由于基于距离度量的说话人分割算法中对窗宽长度的限制，加之没有先验知识，使得两窗语音之间距离的度量欠缺准确性和稳定性。考虑到如果两窗语音属于同一个说话人，那么二者与同一个说话人模型的距离的差异较小，如果两段语音与多个说话人模型的距离的差异都较小则说明二者属于同一个说话人的概率很高，如果两段语音与多个说话人模型的距离的差异较大则说明二者属于同一个说话人的概率较低。使用参考说话人模型作为中介进行距离度量，训练对声学空间充分覆盖的多个参考说话人模型，每个参考说话人模型代表一类说话人的声学特性，利用语音与多个参考说话人模型的距离的差异来度量两段语音之间的距离，并利用性别相关的 UBM 获取语音的性别信息并用来增强距离度量的准确程度，分割后充分利用距离序列上的波峰波谷信息进行误警和漏检的判断。在 NIST SRE 2002 说话人分割数据集合上的实验结果表明，与 BIC、GLR 和 DISTBIC 分割方法比较，FAR 和 MDR 均有明显下降，基于 RSM 的

分割算法相比于传统的 DISTBIC 算法，在新闻采访语音库 BNEWS 漏检率相对下降 34.8%，总错误率（FAR＋MDR）相对下降了 18.7%；在电话交谈语音库 SWBD 上漏检率相对下降 7.5%，总错误率相对下降 10.4%。

2．为了进一步提高说话人分割中距离度量的精度，提出了一种**基于音素识别和文本相关的说话人分割算法**。考虑到在较短语音下（甚至仅有一两个字或词）文本相关的说话人识别可以达到很高的识别性能，要远好于文本无关的说话人识别，其原因在于识别语音和训练语音在内容上的存在着高度一致性或相关性。该算法利用音素识别技术获取语音中的音素信息，对输入语音进行音素识别得到音素序列，对两窗内相同的内容进行文本相关的说话人识别并将其识别结果作为距离度量，通过提升距离度量时内容的一致性和相关性来提升距离度量的准确性，从而改善说话人分割性能。在 TIMIT 数据库上，能够进一步改善距离度量的准确性提高分割性能，相比于 RSM 算法，漏检率相对下降 15.4%，总错误率相对下降 6.8%，相比 GLR 算法，漏检率相对下降 16.9%，错误率相对下降 18.5%。

3．针对说话人检测中的说话人聚类问题，提出一种**基于类纯度约束的说话人聚类算法**。在说话人数目未知的无监督聚类中，如果最终聚类的类数少于说话人的数目，就很有可能使目标说话人的语音被淹没在其他非目标说话人语音中，而如果聚类数多于说话人数目，发生目标说话人语音被淹没的概率会比较小，本文研究了基于类纯度约束的说话人聚类方法。该算法的基本思想是使聚类后达到最短辨认长度要求且类纯度高的有效类数量尽可能多，降低不同说话人的语音被聚到一类内的概率。该算法使用参考说话人模型度量语音间距离，以类内离散度最小、类纯度最大为准则，合并时不仅考虑语音段与类内已有语音的距离，而且考虑合并对于类内离散度的影响，以类内离散度评估类纯度，如果合并对于类内离散度的影响超过阈值，则停止合并而将该段语音单独分类；如果类内语音的总长度达到最短辨认长度要求，则该类单独归为一类不再继续聚类；通过这两点来保证单类的类纯度较高。与传统的 HAC 算法比较，类纯度得到较大提升从而使目标说话人漏检的概率降低；而且时间消耗有较大幅度下降。在 NIST SRE 2006 数据库上，在语音平均长度分别为 2 秒、5 秒和 8 秒的条件下，与传统的 HAC 算法比较，有效类语音比例分别提高了 2.7%、3.8%和 4.6%，目标说话人检测的召回率分别提高了 7.6%、6.2%和 5.1%。

4．针对大规模目标说话人使得说话人辨认的速度降低问题，提出了一种**基于参考说话人和双层结构的说话人快速辨认算法**。目标说话人越多，说话人辨认所需要的时间越长，因此，大规模说话人辨认任务中辨认速度是必须面对的、极其关键的问题。重点研究了目标说话人剪枝算法，目标说话人数量增大

到一定程度后 HSI 和 SMC 算法的辨认性能下降较明显，其根本原因在于其剪枝后保留的目标说话人与聚类中心相似程度高，但不能保证与待辨认语音相似程度高，本文提出了基于参考说话人和双层结构的快速辨认算法，将参考说话人模型组织成双层结构，利用待辨认语音与上层节点之间距离来度量待辨认语音和下层参考说话人之间的相似程度，对目标说话人进行粗剪枝并快速挑选与待辨认语音最相近的一部分下层参考说话人用来度量待辨认语音与目标说话人之间的相似程度。在 CCC VPR-2C2005-6000 数据库上，在 5,200 个目标说话人和 1,000 个集外说话人的开集辨认条件下，在基于 GMM-UBM 架构的说话人辨认系统中，RSMSP 剪枝算法与 SMC 算法相比，运算时间降低 29.3%而辨认准确率提升 1.4%。RSMSP 算法与快速挑选核心分布算法 TBKS 的融合能够进一步提升辨认的运算速度，与 SMC 和 TBKS 的融合算法相比，运算时间降低了 31.9%而辨认准确率提升 1.2%。使用 DISTBIC 分割算法，HAC 聚类算法，SMC+TBKS 快速辨认算法作为说话人检测的基线系统，使用本文提出基于 RSM 的分割算法、基于类纯度约束的聚类算法和基于 RSM 和双层结构的目标说话人剪枝算法的说话人检测系统，较基线系统的目标说话人的召回率提高 6.8%，运算时间降低 31.9%。

6.2　下一步研究的展望

本文对大规模目标说话人检测关键技术进行了一些初步研究，在前人研究的基础上取得了一些研究成果，但同时也发现了一些不足之处。下面将针对这些不足之处，指出今后计划进一步深入开展研究的若干方向。

1．本文提出的基于参考说话人模型的距离度量方法，在训练参考说话人模型时，仅选择了作者认为足够数量的不同说话人，并保持其信道、性别和语言均衡，假设该说话人集合能够覆盖说话人空间，并在该集合中训练代表性的说话人模型作为参考说话人，该集合与复杂的声学空间相比肯定还存在着一定的差距，训练出的参考说话人模型的描述声学特性的能力还需要进一步验证，还需要进一步研究更具推广性的参考说话人训练方法。基于音素识别和文本相关的说话人分割方法，受到音素识别性能，两窗语音中是否存在相同音素的制约，如何进一步寻找两窗语音间的共同内容尤其是不依赖于内容的隐藏信息并准确的获取这些隐藏信息对提高分割性能很有帮助。在现实中，人对说话人转换的判断能力是相当强大的，这是因为人综合运用了多种信息，如说话人性别、音色、口音、韵律以及语音环境变化等。如何准确

提取出更多的区分性强的信息，并与已有的特征和方法进行融合，是提高说话人分割性能的途径之一，也是今后的一个研究方向和内容[132-138]。

2．本文研究的说话人聚类算法仅仅是在前人研究的基础上做了一点点探索，算法还受到一定的限定条件的制约，只是针对说话人检测的应用有一定效果，在其他说话人聚类应用场景中并不一定能够适用，对于语音中存在的同时发音的情况本文并未过多考虑。如何准确判断出一段多说话人语音中的说话人数目还是今后的一项重要研究内容。类纯度对说话人辨认的影响非常大，多个说话人同时发音对聚类的类纯度影响很大，如果能够挑选出多人同时发音数据是对于进一步提高聚类的类纯度是有帮助的，是今后的一个重要的研究内容。

3．由于现有数据条件的限制，本文的快速辨认算法的实验在 6,200 人数量级的数据库中进行，对于实际应用中的几万甚至百万级的目标说话人规模，本文的方法还需要进一步验证。如何在待辨认语音长度较短的条件下提高说话人辨认性能的也是今后一项重要的研究内容，这可以降低大规模目标说话人检测对于分割、聚类性能的依赖程度。

参考文献

[1] S. Pruzansky. Pattern-matching procedure for automatic talker recognition. Journal of the Acoustical Society of America, 1963, 35(3):354-358.

[2] J. Campbell. Speaker recognition: a tutorial. In Proceedings of IEEE, 1997, 85(9):1437-1462.

[3] M. Kotti, V.Moschou, C. Kotropoulos. Speaker segmentation and clustering. Signal Processing, 2008, 88:1091-1124.

[4] V. R. Apsingekar and P. Leon. Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications. IEEE Transactions on Audio Speech and Language Processing, 2009, 17(4):848-853.

[5] NIST SRE, Online available, http://www.itl.nist.gov/iad/mig/tests/sre/.

[6] NIST RT, Online available, http://www.itl.nist.gov/iad/mig/tests/rt/.

[7] D. Moraru, S. Meignier, L. Besacier, *et al*. The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2003, 2:89-92.

[8] D. Moraru, S. Meignier, C. Fredouille, *et al*. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2004:373-376.

[9] D. Moraru, L. Besacier, S. Meignier, C. Fredouille, J.-F. Bonastre. Speaker diarization in the ELISA consortium over the last years. In Proceedings of 2004 RT-04 Fall Workshop, 2004.

[10] C. Barras, X. Zhu, S. Meignier, J.L. Gauvain. Multistage speaker diarization of broadcast news. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(5):1505-1512.

[11] X. Zhu, C. Barras, L. Lamel and J. L. Gauvain. Multi-Stage Speaker Diarization for Conference and Lecture Meetings. In Rich Transcription 2007 Meeting Recognition Workshop, 2007.

[12] P. Sivakumaran, J. Fortuna and A. M. Ariyaeeinia. On the use of the Bayesian information criterion in multiple speaker detection. In Proceedings of European Conference on Speech Communication and Technology, Eurospeech, 2001:795-798.

[13] S. S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In DARPA Speech Recognition Workshop, 1998.

[14] M. Kotti, E. Benetos, C. Kotropoulos. Automatic speaker change detection with the Bayesian information criterion using MPEG-7 features and a fusion scheme. In Proceedings of the 2006 IEEE International Symposium on Circuits and Systems, 2006:856-1859.

[15] A. Tritschler, R. Gopinath. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In Proceedings of the 6th European Conference on Speech Communication and Technology, Budapest, Hungary, September 1999:679-682.

[16] J. Ajmera, I. McCowan, H. Bourlard. Robust speaker change detection. IEEE Signal Processing Letters, 2004, 11(8):649-651.

[17] H. Gish, M. H. Siu and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. IEEE International Conference on Acoustics Speech and Signal Processing, 1991:873-876.

[18] H. Gish and M. Schmidt. Text-independent speaker identification. IEEE Signal Processing Mag. 1994, 11:18-32.

[19] S. Meignier, D. Moraru, C. Fredouille, J. F. Bonastre, L. Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. Comput. Speech Language, 2006, 20 (2-3):303-330.

[20] S. Know, S. Narayanan. Unsupervised speaker indexing using generic models. IEEE Transactions on Speech Audio Process, 2005, 13 (5):1004-1013.

[21] M. A. Siegler, U. Jain, B. Raj and R. M. Stern. Automatic segmentation classification and clustering of broadcast news audio. In DARPA Speech Recognition Workshop, 1997:97-99.

[22] L. Lu, H. Zhang. Unsupervised speaker segmentation and tracking in real-time audio content analysis. Multimedia Systems, 2005, 10 (4):332-343.

[23] P. Delacourt and C. J. Wellekens. DISTBIC: a speaker-based segmentation for audio data indexing. Speech Communication, 2000, 32(1-2):111-126.

[24] H. Harb, L. Chen. Audio-based description and structuring of videos. International Journal Digital Libraries, 2006, 6 (1):70-81.

[25] D. A. Reynolds, E. Singer, B. A. Carlson, J. J. McLaughlin, *et al*. Blind clustering of speech utterances based on speaker and language characteristics. In Proceedings of International Conference on Spoken Language Processing, ICSLP, 1998:610-613.

[26] S. Meignier S, J. F. Bonastre and I. Magrin-Chagnolleau. Speaker utterances tying among speaker segmented audio documents using hierarchical classification: towards speaker indexing of audio databases. In Proceedings of International Conference on Spoken Language Processing, ICSLP, 2002, 1:573-576.

[27] S. Know, S. Narayanan. Speaker change detection using a new weighted distance measure. In Proceedings of the International Conference on Spoken Language Processing, ICSLP, 2002, 4:2537-2540.

[28] M. C. Ivan, E. R. Aaron and S. Parthasarathy. Detection of target speakers in audio databases. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 1999:821-824.

[29] S. Meignier, J. F. Bonastre and S. Igounet. E-HMM approach for learning and adapting sound models for speaker indexing. In 2001: A Speaker Odyssey, Chania, Crete, 2001:175-180.

[30] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 1995, 3(1):72-83.

[31] S. Galliano, E. Geoffrois, G. Gravier, *et al*. Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In Proceedings of LREC, Genoa, 2006:315-320.

[32] D. Graff, Z.Wu, R. MacIntyre and M. Liberman. The 1996 broadcast news speech and language-model corpus. In Proceedings of the 1997 DARPA Speech Recognition Workshop, 1997.

[33] C. Fredouille, D. Moraru, S. Meignier, *et al*. The NIST 2004 spring rich transcription evaluation: two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation. In RT2004 Spring Meeting Recognition Workshop, 2004:5-8.

[34] T. Wu, L. Lu, K. Chen, H. Zhang. UBM-based real-time speaker segmentation for broadcasting news. In Proceedings of IEEE International Conference Acoustics Speech and Signal Processing, ICASSP, 2003:193-196.

[35] T. Wu, L. Lu, K. Chen, H. Zhang. Universal background models for real-time speaker change detection. In Proceedings of the 9th International Conference on Multimedia Modeling, 2003:135-149.

[36] G.-X. Yi, Q.-Y. Li, Zh.-S. Lin, X.-H. Wu, H.-Sh. Chi. Speaker segmentation based on model scoring. Technical Acoustics, 2005, (24):218-221.

[37] 白俊梅，张树武，徐波. 广播电视中的目标说话人跟踪技术. 声学技术，2005，(24):34-38.

[38] 吕萍，颜永红. 广播新闻语料自动识别系统. 声学技术，2005，(24):109-112.

[39] 杨旻. 多层次说话人分割及相关算法研究[硕士学位论文]. 浙江大学，2006.

[40] M. Collet, D. Charlet, F. Bimbot. A correlation metric for speaker tracking using anchor models. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 2003, 1:713-716.

[41] A. Higgins. YOHO speaker verification. Presented at the Speech Research Symposium. Baltimore, MD, 1990.

[42] B. S. Atal. Automatic recognition of speakers from their voices. Proc. IEEE, 1976, 64(4):460-475.

[43] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustic, Speech and Signal Processing, 1980, 28:357-366.

[44] 甄斌，吴玺宏，刘志敏，迟惠生. 语音识别和说话人识别中各倒谱分量的相对重要性. 北京大学学报（自然科学版），2001，37(3):371-378.

[45] S. Furui. Comparison of speaker recognition methods using static features and dynamic features. IEEE Transaction on Acoustics, Speech, and Signal Processing, 1981, 29(3):342-350.

[46] R. Sinha, S. Tranter, M. Gales, P. Woodland. The Cambridge University March 2005 speaker diarization system. In Proceedings of the European Conference on Speech Communication and Technology, 2005:2437-2440.

[47] H. G. Kim, T. Sikora. Comparison of MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, 5:925-928.

[48] H. Hermansky. Perceptual linear prediction (PLP) analysis for speech. Journal of the Acoustic Society of America, JASA, 1990, 87(4):1738-1752.

[49] S. E. Tranter, K. Yu, G. Evermann, P. C. Woodland. Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2004:433-437.

[50] L. Lu, H. Zhang. Speaker change detection and tracking in real-time news broadcast analysis. In Proceedings of the ACM Multimedia, 2002:602-610.

[51] D. Wang, L. Lu, H.-J. Zhang. Speech segmentation without speech recognition. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2003, 1:468-471.

[52] L. Lu, H.-J. Zhang and H. Jiang. Content analysis for audio classification and segmentation. IEEE Transactions on Speech and Audio Processing, 2002, 7(10):504-516.

[53] A. Adami, R. Mihaescu, D. A. Reynolds and J. Godfrey. Modeling prosodic dynamics for speaker recognition. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2003:106-110.

[54] B. Peskin, J. Navratil, J. Abramson, *et al*. Using prosodic and conversational features for high-performance speaker recognition. Report from JHU WS'02, ICASSP, 2003:792-795.

[55] G. Friedland, O. Vinyals, Y. Huang, C. Müller. Prosodic and other Long-Term Features for Speaker Diarization. In Proceedings of International Conference on Acoustics Speech and Signal Processing, ICASSP, 2009:4077-4080.

[56] G. Friedland, O. Vinyals, Y. Huang, C. Müller. Fusing short term and long term features for improved speaker diarization. IEEE transactions on Audio Speech and Language Processing, 2009, 17 (5):985-993.

[57] H.-W. Sun, T. L. Nwe, B. Ma and H.-Z. Li. Speaker Diarization for Meeting Room Audio. The 10th Annual Conference of the International Speech Communication Association, InterSpeech, 2009:900-903.

[58] W. Wu, T. F. Zheng, M.-X. Xu and H.-J. Bao. Study on Speaker Verification on Emotional Speech. In Proceedings of International Conference on Spoken Language Processing, Interspeech, 2006:2102-2105.

[59] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, *et al*. Speaker recognition general classifier approaches and data fusion methods. Pattern Recognition, 2002, 35(12):2801-2821.

[60] 杨行峻，迟惠生，等.《语音信号数字处理》．电子工业出版社，1995.

[61] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustic, Speech and Signal Processing, 1978, 26(1):43-49.

[62] A. L. Higgins, L. G. Bahler and J. E. Porter. Voice identification using nearest neighbor distance measure. In Proceedings IEEE International Conference on Acoustics, Speech, Signal Processing, 1993:375-378.

[63] J. Hertz, A. Krogh and R. G. Palmer. Introduction to the theory of neural computation.Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, Reading, Mass, USA. 1991.

[64] S. Haykin. Neural networks: a comprehensive foundation. Macmillan, New York, NY, USA, 1994.

[65] V. N. Vapnik. The nature of statistical learning theory. Springer-Verlag, New York, 1995.

[66] M. Schmidt and H. Gish. Speaker identification via support vector classifiers. In Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, 1996, 1:105-108.

[67] Y. Gu and T. Thomas. A text-independent speaker verification system using support vector machines classifier. In Proceedings of European Conference on Speech Communication and Technology, Eurospeech, 2001:1765-1769.

[68] D. Xin, Z.-H. Wu and Y.-Ch Yang. Exploiting support vector machines in hidden Markov models for speaker verification. In Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP, 2002:1329-1332.

[69] D. A. Reynolds, T. F. Quatieri and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 2000, 10:19-41.

[70] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, ICASSP, 2006:97-100.

[71] V. Wan, W. M. Campbell. Support vector machines for speaker verification and identification. In Proceedings of the Neural Networks for Signal Processing, 2000, 10:775-784.

[72] J. Kharroubi, D. D. Petrovska and G. Chollet. Combining GMMs with support vector machines for textindependent speaker verification. In Proceedings European Conference on Speech Communication and Technology, Eurospeech, Aalborg, Denmark, September 2001:1757-1760.

[73] P. Kenny, G. Boulianne and P. Dumouchel. Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing, 2005, 13(3):345-354.

[74] P. Kenny, P. Ouellet, N. Dehak *et al*. A Study of Inter-Speaker Variability in Speaker Verification. IEEE Transactions on Audio Speech and Language Processing, 2008, 07:980-988.

[75] S.-C. Yin, R. Rose, P. Kenny and P. Dumouchel. A Joint factor analysis approach to progressive model adaptation in text independent speaker verification. IEEE Transactions on Audio Speech and Language Processing, 2007, 15 (7):1999-2010.

[76] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio Speech and Language Processing, 2007, 15 (4):1448-1460.

[77] R. Vogt and S. Sridharan. Experiments in session variability modeling for speaker verification. In Proceedings of International Acoustic, Signal, Speech and Processing, 2006:897-900.

[78] C.-H. Lee, C.-H. Lin and B.-H. Juang. A study on speaker adaptation of parameters of continuous density hidden Markov models. IEEE Transactions on Acoustic and Speech Signal Processing, 1991, 39(4):806-814.

[79] C.-H. Lee and J. L. Gauvain. Speaker adaptation based on MAP estimation of HMM parameters. In Proceedings International Conference on Acoustics, Speech, and Signal Processing, 1993, 2:652-655.

[80] J. L. Gauvain, and C. H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process, 2, 1994:291-298.

[81] 李虎生，刘加，刘润生. 语音识别说话人自适应研究现状及发展趋势. 电子学报，2003，31(1):103-108.

[82] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language. 1995, 9:171-185.

[83] C. J. Leggetter and P. C. Woodland. Flexible speaker adaptation for large vocabulary speech recognition. In Proceedings of European Conference on Speech Communication and Technology, Eurospeech, 1995:1155-1158.

[84] C. J. Leggetter. Improved acoustic modeling for HMMs using linear transformations. PhD thesis. Cambridge University, 1995.

[85] M. Liu, Eric Chang, Beiqian Dai. Hierarchical Gaussian mixture model for speaker verification. In Proceedings of International Conference on Spoken Language Processing, ICSLP, 2002:1353-1356.

[86] G. Wang, X.-J. Wu, T. F. Zheng, L.-L. Wang and C.H. Zhang. Regression-class Tree based Method for Efficient Speaker Identification. In Proceedings Asia-Pacific Signal and Information Processing Association-Annual Summit and Conference, APSIPA ASC, Sapporo Japan, 2009, 2010:462-465.

[87] Z.-Y. Xiong, T. F. Zheng, Z.-J. Song, F. Soong, W.-H. Wu. A tree-based kernel selection approach to efficient Gaussian mixture model–universal background model based speaker identification. Speech Communication, 2006, 48:1273-1282.

[88] R. Auckenthaler and J. Mason. Gaussian selection applied to text-independent speaker verification. In Proceedings of Speaker Odyssey-Speaker Recognition Workshop, 2001.

[89] B. Xiang and T. Berger. Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. IEEE Trans. Speech Audio Process. 2003, 11(5):447-456.

[90] 熊振宇，郑方，宋战江，吴文虎. 基于树形通用背景模型的高效说话人辨认. 清华大学学报（自然科学版）, 2006, Vol. 46, No. 7, pp:1305-1308.

[91] R. Saeidi, T. Kinnunen, H. Mohammadi, R. Rodman, *et al*. Joint Frame and Gaussian Selection for Text Independent Speaker Verification. ICASSP, 2010:4530-4533.

[92] J. McLaughlin, D. A. Reynolds and T. Gleason. A study of computation speed-ups of the GMM-UBM speaker recognition system. In Proc. 6th European Conference on Speech Communication and Technology, Eurospeech, Budapest, Hungary, 1999:1215-1218.

[93] Pellom and J. Hansen. An efficient scoring algorithm for gaussian mixture model based speaker identification. IEEE Signal Processing Letters, 1998, 5(11):281-284.

[94] Z.-Y. Xiong, T. F. Zheng, Z. j. Song, W. h. Wu. Combining Selection Tree with Observation Reordering Pruning for Efficient Speaker Identification Using GMM-UBM. In Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, ICASSP, 2005:625-628.

[95] T. Kinnunen, E. Karpov, and P. Franti. Real-time speaker identification and verification. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(1):277-288.

[96] B. Sun, W. Liu and Q. Zhong. Hierarchical speaker identification using speaker clustering. In International Conference on Natural Language and Knowledge Engineering, 2003:299-304.

[97] 边肇祺，张学工等. 模式识别. 清华大学出版社, 2000.

[98] A. V. Hall. Methods for demonstrating resemblance in taxonomy and ecology. Nature, 1967, 214: 830-831.

[99] X.-D. Huang, A. Acero, H.-W. Hon. Spoken Language Processing: a Guide to Theory, Algorithm, and System Development. 2001, Prentice-Hall.

[100] Chinese Corpus Consortium. Online available, http://www. CCCForum.org/.

[101] J. Rissanen. Stochastic complexity in statistical inquiry. Series in Computer Science, 1989, Vol. 15. World Scientific, Singapore, Chapter 3.

[102] P. Sivakumaran, A. M. Ariyaeeinia and J. Fortuna. An effective unsupervised scheme for multiple-speaker-change detection. In Proceedings of International Conference on Spoken Language Processing, ICSLP, 2002:569-572.

[103] J. F. Bonastre, P. Delacourt, C. Fredouille, *et al*. A speaker tracking system based on speaker turn detection for NIST evaluation. In Proceedings of International Conference on Acoustics Speech and Signal Processing, ICASSP, 2000:1177-1180.

[104] D. Liu and F. Kubala. Fast speaker change detection for broadcast news transcription and indexing. In Proceedings of European Conference on Speech Communication and Technology, Eurospeech, 1999:1031-1034.

[105] A. Dempster, N. Laird and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Stat. Soc. 1977, 39:1-38.

[106] http://www.itl.nist.gov/iad/mig/tests/sre/2002/2002-spkrec-evalplan-v60.pdf.

[107] C. Boehm and F. Pernkopf. Effective metric-based speaker segmentation in the frequency domain. In Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2009:4081-4084.

[108] P. Matejka, P. Schwarz, J. Cernocky, P. Chytil. Phonotactic language identification using high quality phoneme recognition. In Proceedings of the 9th European Conference on Speech Communication and Technology, Interspeech 2005-Eurospeech, 2005:2237-2240.

[109] D. A. Reynolds. An Overview of Automatic Speaker Recognition Technology. In Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing ICASSP, 2002:4072-4075.

[110] W. D. Andrews, M. A. Kohler and J. P. Campbell. Phonetic Speaker Recognition, In Proceedings of European Conference on Speech Communication and Technology, Eurospeech, 2001:2517-2520.

[111] 秦兵，陈惠鹏，李光琪，刘松波. 文本有关的话者确认系统. 哈尔滨工业大学学报，2000，32(4):16-18.

[112] 邓浩江，杜利民，万洪杰. 似然得分归一化及其在文本无关说话人确认中的应用. 电子与信息学报，2005，27(7):1025-1029.

[113] J. S. Garofolo. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, 1993.

[114] A. K. Jain, M. N. Murty, P. J. Flynn. Data clustering: A review. The Annals of Statistics 6 (2), 1978:461-464.

[115] H. Jin, F. Kubala and R. Schwartz. Automatic speaker clustering. DARPA Speech Recognition Workshop, 1997.

[116] A. Solomonoff, A. Mielke, M. Schmidt and H. Gish. Clustering speakers by their voices. In Proceedings of International Conference on Acoustics Speech and Signal Processing ICASSP, 1998:757-760.

[117] J. Ajmera, H. Bourlard, I. Lapidot, I. McCowan. Unknown multiple speaker clustering using HMM. In Proceedings of the International Conference on Spoken Language Processing, ICSLP, 2002:573-576.

[118] S. E. Johnson, P. C. Woodland. Speaker clustering using direct maximization of the MLLR-adapted likelihood. In Proceedings of the International Conference on Spoken Language Processing, ICSLP, 1998, 5:1775-1778.

[119] R. Faltlhauser and G. Ruske. Robust speaker clustering in eigenspace. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2001.

[120] Y. Moh, P. Nguyen, J. C. Junqua. Towards domain independent speaker clustering. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003:85-88.

[121] D. Liu, F. Kubala. Online speaker clustering. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2004, 1:333-336.

[122] 王炜, 吕萍, 颜永红. 一种改进的基于层次聚类的说话人自动聚类算法. 声学学报, 2008, 33(1):9-14.

[123] W. M. Rand, Objective criteria for the evaluation of clustering methods. Journal American Stat. Assoc. 1971, 66:846-850.

[124] W.-H. Tsai and H.-M. Wang. Speaker clustering of unknown utterances based on maximum purity estimation. In Proceedings of the European Conference on Speech Communication and Technology, Interspeech, 2005:3069-3072.

[125] W.-H. Tsai and H.-M. Wang. Evolutionary Minimization of the Rand Index for Speaker Clustering. Computer, Speech and Language, 2009, 23:165-175.

[126] K. J. Han and S. S. Narayanan. A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system. In Proceedings of the 8th Annual Conference of the International Speech Communication Association, Interspeech, 2007:1853-1856.

[127] K. J. Han, S. Kim and S. S. Narayanan. Robust speaker clustering strategies to data source variation for improved speaker diarization. In Proceedings of Automaic Speech Recognition and Understanding, ASRU, 2007:262-267.

[128] X. Anguera, C. Wooters and J. Hernando. Purity algorithms for speaker diarization of meetings data. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006:1025-1028.

[129] C. Jankowski, A. Kalyanswamy, S. Basson, J. Spitz. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990, 1:109-112.

[130] K. P. Li and J. E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, ICASSP, 1988:595-598.

[131] 熊振宇. 大规模、开集、文本无关说话人辨识研究[博士学位论文]. 北京： 清华大学计算机系, 2005.

[132] D. Reynolds, W. Andrews, J. Campbell, *et al*. The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition. In Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, ICASSP, 2003:784-787.

[133] 邓菁. 电话信道下多说话人识别研究[博士学位论文]. 北京： 清华大学计算机系, 2007.

[134] L. R. Rabiner and B.-H. Juang. Fundamentals of speech recognition. Signal Processing. Prentice-Hall, NJ, 1993.

[135] Z. H. Chen, Y. F. Liao Y F and Y. T. Juang. Prosody modeling and eigen-prosody analysis for robust speaker recognition. In Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, ICASSP, Philadelphia, USA, 2005:185-188.

[136] Adami A G. Prosodic modeling for speaker recognition based on sub-band energy temporal trajectories. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Philadelphia, USA, 2005:189-192.

[137] A. Mijail, A. Anil, Z. Philip, *et al*. A Bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition. In Proceedings of the 9th European Conference on Speech Communication and Technology, Interspeech'2005-Eurospeech, 2005:2009-2013.

[138] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen. Maximum likelihood discriminant feature spaces, In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2000, 2:1129-1132.

致　谢

衷心感谢我的导师郑方研究员四年来对我的悉心指导和关怀。郑老师严谨求实，平易近人，无论在学业上还是在生活上都给予了我莫大的关心和帮助。在我的研究遇到困难而彷徨无助的时候，和郑老师的讨论总能给我以极大的启发。郑老师的言传身教将使我受益终生。在此，谨向恩师致以最诚挚的谢意！

感谢语音与语言技术中心的其他老师，包括方棣棠教授、徐明星副教授、邬晓钧老师；感谢实验室的张利鹏、王琳琳、罗灿华、张陈昊等同学。他们与我进行了许多有益的讨论，同时给予了我很多学习与工作上的支持和帮助，在此一并向他们表示感谢。

最后，衷心感谢我的家人和朋友，他们无私的爱和默默的关怀，一直伴随着我的奋斗过程。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1976 年 9 月出生于河北省秦皇岛市。

1995 年 9 月考入国防科学技术大学机械电子工程与仪器系，1999 年 7 月本科毕业并获得工学学士学位。

1999 年 9 月考入国防科学技术大学机械电子工程与仪器系攻读硕士，2002 年 3 月硕士毕业获得工学硕士学位。

2002 年 3 月进入空军第五研究所工作，2004 年 9 月进入空军装备研究院地面防空装备研究所工作。

2007 年 9 月考入清华大学计算机系攻读博士至今。

发表的学术论文

[1] Gang Wang, Xiaojun Wu and Thomas Fang Zheng, Linlin Wang and Chenhao Zhang. Regression-class Tree based Method for Efficient Speaker Identification. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, Singapore, 2010, pp: 462-465. (To be EI indexed)

[2] Gang Wang, Xiaojun Wu and Thomas Fang Zheng. Using Phoneme Recognition and Text-dependent Speaker Verification to Improve Speaker Segmentation for Chinese Speech. Interspeech, Makuhari, Chiba, Japan, 2010, pp: 1457-1460. (To be EI indexed)

[3] Gang Wang and Thomas Fang Zheng. Speaker segmentation based on between-window correlation over speakers' characteristics. In Proceedings of Asia-Pacific Signal and Information Processing Association-Annual Summit and Conference, APSIPA ASC, Sapporo, Japan, 2009. (EI indexed, Accession Number: 20101512838302)

[4] Gang Wang and Thomas Fang Zheng. Using MMSE to improve session variability estimation. International Journal Biometrics, 2010, Vol. 2, No. 4, pp: 350-357.

[5] 王刚，郑方. 电话信道下应用 DMFCC 进行说话人识别的研究. 清华学报（自然科学版）. 2009, Vol. 49, No. 10, pp: 1597-1600.（EI indexed, Accession number: 20094612457033）

研究成果

[1] 中国建设银行 95533 声纹电话银行项目，主要工作有 UBM 数据的选取与训练，防止录音假冒的流程设计，开发和测试工作，已经通过验收并上线运行（合作完成）。

[2] 公安部“十一五”海量目标说话人筛选系统的设计开发工作，已通过公安部验收并交付使用（合作完成）。