说话人识别中的时变鲁棒性问题研究

（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 琳 琳

指导教师：郑　　方　研究员

二〇一三年四月

Research on Time-Varying Robustness in Speaker Recognition

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Wang Linlin

Dissertation Supervisor: Professor Zheng Fang

April, 2013

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

本文对人类的声纹中普遍存在的时变现象进行了深入研究，主要工作包括：

1. 建立了一个适合于声纹时变性研究的声纹数据库 Chronos。在综合分析了现有的时变声纹数据库的基础上，提出了时变声纹数据库的总体设计原则：“尽最大可能保证时间是唯一变化的因素”。采用固定语料作为录音提示文本以尽量减少语音内容差异带来的影响，要求说话人以朗读方式发音以尽量减少说话方式变化带来的影响。数据库录制采用梯度的时间间隔，即“最初语音采集比较频繁，之后间隔越来越长”，既保证不同时间间隔的覆盖性，又可减少录音成本。

2. 提出了说话人确认系统时变鲁棒性的综合评价准则。对于时变说话人确认系统而言，每次录音会话均存在一个等错误率；等错误率的均值代表了系统的平均性能水平，而标准差则代表了系统性能随时间的变化性，因此本文以这一系列会话的等错误率的均值和标准差作为衡量系统的时变鲁棒性的重要评价指标，并定义两者的乘积为时变鲁棒性的综合评价准则。

3. 提出了基于以 F-ratio 为中间准则计算频带区分度的时变鲁棒特征提取算法。提出了频带整体区分度的概念，验证了不同频带对于时变说话人识别具有不同区分度的假设。采用基于频带能量的 F-ratio 为中间准则来计算各个频带的整体区分度。探讨了两种提高时变鲁棒性的特征提取方法：在传统 MFCC 计算过程中的滤波之前进行频率弯折以及滤波之后对滤波器的输出进行加权。前者是通过增加或减少滤波器的数目（调整滤波器分辨率）来强调或弱化相应频段，而后者则是通过直接对滤波器的输出进行加权来强调或弱化相应频段。两种方法相对于 MFCC 特征，说话人确认的时变鲁棒性综合评价指标分别提升 26.90%和 5.45%。

4. 提出了基于性能驱动的频率弯折方法的特征提取算法。本算法以性能（时变鲁棒性综合评价准则）作为目标函数去优化频带的整体区分度：在保持其他所有频带分辨率不变的前提下，单独强调某一指定频带所对应的系统性能作为该频带的整体区分度指标。据此对各频带分别进行相应的频率弯折，得到性能驱动的声纹特征。该特征可将说话人确认的时变鲁棒性综合评价指标提升 32.47%。

5. 提出了基于区分性训练的滤波器输出加权方法的特征提取算法。本算法也是性能驱动的方法，利用具有区分性的特征提取的思想，给定滤波器输出权重一个初始序列，经过建模和打分过程，依据系统反馈的性能、通过根据时变鲁棒性综合评价指标而提出的 MCE*MSV 准则来自动调整输出权重，如此反复迭代，直到收敛到一个性能优化的权重序列。据此对各频带对应的滤波器输出进行加权，得到声纹特征。该特征可将说话人确认的时变鲁棒性综合评价指标提升 34.08%。

关键词：说话人识别；时变现象；时变鲁棒

Abstract

The focus of this dissertation is the time-varying issue in speaker recognition and the time-varying robustness is explored. Major efforts and contributions are:

1. A proper longitudinal voiceprint database that specially focuses on the time-varying issue. After analyzing existing speech databases with the time-varying attribute, we designed to create a fixed-text read speech database with 16 recording sessions within a time span of 3 years. Since the time-varying effect was the only focus, other factors, such as recording equipment, software, conditions and environment were kept as constant as possible throughout all recording sessions. Gradient time intervals were used, with the length of intervals increasing gradually.

2. Performance evaluation index for a time-varying speaker recognition system. For a time-varying speaker verification task, there are generally a series of EERs, corresponding to each recording session. Then when comparing the performance of two systems, we are indeed comparing two arrays of EERs. Therefore, it is natural to use mean and standard deviation of each array of EERs to evaluate the overall performance of a system. The mean value serves as an indicator of the averages performance of sessions, while the standard deviation value serves as an indicator of the time-varing robustness across sessions. Specifically in this paper, the product of those two values is used to evaluate the overall time-varying speaker verification performance.

3. Time-varying robust feature extraction algorithms with discrimination sensitivity of frequency bands calculated through F-ratio. The concept of overall discrimination sensitivity of frequency bands regarding the time-varying speake recognition task was proposed. Efforts were made to identify frequency bands that revealed high discrimination sensitivity for speaker-specific information, while low discrimination sensitivity for time-varying session-specific information. F-ratio was employed as an intermediary criterion to calculate the overall discrimination sensitivity based on the log-energy spectrum. Thus according to the overall discrimination sensitivity, tme-varying robust feature extraction algorithms were presented during feature extraction of cepstral coefficients with different emphasis on different frequency bands from two aspects: pre-filtering frequency-warping and post-filtering filter-bank outputs weighting. Experimental results showed that the two algorithms outperformed the baseline MFCC by 26.90% and 5.45%, respectively.

4. Performance-driven feature extraction algorithm based on frequency warping. This algorithm evaluated the overall discrimination sensitivity of frequency bands from a performance-driven point of view instead of the F-ratio criterion. Specifically, the overall discrimination sensitivity of a designated frequency band is determined by the overall performance of a time-varying speaker recognition system, which made use of frequency-warping approach to soly emphasize the designated frequency band, leaving other unchanged. Finally, frequency warping was performed and experimental results showed that it yielded a better result than MFCC, with a gain of 32.47 in overall performance.

5. Discriminative feature extraction algorithm based on filter-bank outputs weighting. This was also a performance-driven approach, yet it was designed for the filter-bank outputs weighting method. After resigning an initial series of weights for filter-bank outputs, speaker modeling and utterance scoring were performed; then according to the performance feedback, the series of weights were adjusted by the proposed MCE*MSV criterion. After several iterations of such a process, the best series of weights were found automatically. The MCE*MSV criterion was proposed to minimize the target optimization function of the error rates of recording sessions and their standard deviation. The best series of weights were applied to filter-bank outputs and experimental results showed that it worked better than MFCC by 34.08%.

Key words: speaker recognition; time-varying issue; time-varying robustness

目 录

第1章　绪论

1.1　说话人识别应用背景

1.1.1　说话人识别技术概述

说话人识别（Speaker Recognition），又称声纹识别（Voiceprint Recognition），它是利用语音中所含有的说话人的个性信息来自动识别话者身份的一种生物认证技术（Pruzansky, 1963; Campbell, 1997）。说话人识别本质上是模式识别问题的一种，因此一个典型的说话人识别系统一般由训练（将用户预留语音训练成为说话人模型 ，也称声纹预留）和识别（判断一个说话人未知语音是否来自指定说话人，也称声纹验证）两个阶段（或者部分）构成。

根据应用的范畴可分为说话人辨认（Speaker Identification）和说话人确认（Speaker Verification）两类（Campbell, 1997）。顾名思义，辨认指的是判定待识别的一段语音属于候选说话人集合（又称目标说话人集合）中的某一位的技术，其辨认结果为候选说话人集合中哪一位最有可能是目标说话人，它是一个多选一的问题；而确认指的是确定待识别的一段语音是否由其所声明的目标说话人发生的技术，其确认结果为“是”该目标说话人（接受）或“不是（否）”该目标说话人（拒绝）的语音，它是一个二选一的问题。其中说话人辨认又可详细区分为闭集（Close-set）识别和开集识别（Open-set）两类。所谓闭集识别，是指待识别的这段语音必定属于候选说话人集合中的某一位，即待识别语音为集内说话人；所谓开集识别，是指待识别的这段语音有可能不属于候选说话人集合中的任何一位，即待识别语音可能为集外说话人，于是系统存在拒识的情况。

根据待识别语音的文本内容可分为文本无关（Text-independent）和文本相关（Text-dependent）两类（Campbell, 1997）。文本无关指的是说话人识别系统对于语音文本内容无要求，即无论训练还是识别过程，用户均可随意说出一段有效语音长度足够的话；而文本相关指的是说话人识别系统要求用户必须按照事先指定的文本内容进行发音，训练和识别过程对用户要求更高。一般说来，文本相关的说话人识别系统性能要相对好于文本无关的情况；当然，文本无关的说话人识别系统，使用更方便，其应用灵活性要远好于文本相关的情况。

与其他的生物认证技术（如：指纹识别、虹膜识别、脸部识别、掌纹识别等等）相比，语音作为目前人类最自然、最方便和最有效的一种交流方式，其应用

有着天然的优势：说话人识别技术对于语音的采集并不涉及到敏感的用户个人隐私，容易为公众所接受；同时它是一种非接触识别技术，易于依赖已有的电话、网络等资源进行远程应用推广，甚至在某些应用场景中，说话人的语音是唯一可以轻易获取到的生物特征。

1.1.2　说话人识别技术应用

说话人识别技术在商业领域应用广泛，可以通过语音为多种商业服务提供访问控制，包括语音拨号、电话银行服务、电话购物、数据库访问服务、信息查询和预订服务、语音邮件、机密信息的安全控制，以及计算机的远程访问等（Furui, 1997）。2006年荷兰最大的银行荷银ABN AMRO，利用美国Voice Vault公司所开发的说话人识别系统，并结合事先设定的隐私问题，率先在高隐蔽性的银行产业中成功应用了此项技术。2009年，澳大利亚国家银行NAB也在其电话银行中使用了Telstra & Salmat VeCommerce公司提供的VeSecure说话人识别技术。2011年中国建设银行在电话银行部分业务中使用了声纹识别技术进行客户身份的确认，并正在试点推广之中。

除了商业领域的实际应用，说话人识别技术在公共服务领域也有重要的应用价值。美国健康保险供应商Wellpoint公司利用这一技术，生成具有法律约束力的数字签名，这样就可以远程使用语音来“签名”，而不用等文件寄达后再签署。说话人识别也可用于社会保险行业，例如远程确定投保人的生存状况等，目前也是热点应用之一。

此外，公共安全和国防安全领域更是长期关注说话人识别技术及应用的重要领域（Kunzel, 1994）。通过电话侦听（或其他方式）采集到的语音进行自动的身份辨认，对于各种电话勒索、绑架、追逃、电话人身攻击等等案件或者国防情报侦查，在一段录音中查找出目标说话人（或嫌疑人）或者缩小侦查范围。从海量的语音信息中提取目标说话人的语音，可以很好地消除判别语音身份过程中可能出现的人为误差，为侦查提供可靠的情报，提高工作效率。

而应用的推广从来都是与技术本身的发展进步相辅相成的。近年来，随着说话人识别技术的发展和逐步成熟，在限定条件下说话人识别已可获得令人满意的效果。但在实际应用中，声纹预留与声纹验证条件的不匹配是制约一个实际的说话人识别系统性能的重要因素：说话人识别技术面临着诸如背景噪声、信道差异、多说话人以及短语音等挑战，而探索如何更好地消除背景噪声、克服信道失配、多说话人分割聚类以及在短语音条件下进行说话人识别研究，也一直是国内外的研究热点和重点。而其中往往被忽视的一个现象是，在如上所述说话人识别技术

的各方面典型应用之中，声纹预留和声纹验证之间往往相隔一段时间，而随时间变化个体声纹是否依然保持稳定就成为了说话人识别系统走向实用所无法回避的一个问题。本课题研究即基于此。

1.2　说话人识别中的时变问题

“声纹（Voiceprint）”这个概念从诞生之初起就一直伴随着其是否随时间而变化的质疑。Lawrence G. Kersta于1962年在《自然》杂志上发表《声纹识别》一文指出：同指纹一样，人的语音中也存在着身份相关的可识别的唯一性（Kersta, 1962）。研究人员利用复杂的电动机械设备所产生的语谱图证实了这种唯一性的存在，并且在基于视觉对比的识别算法下使得识别率达到了99.65%。尽管受指纹的启发，他们将语音中所含有的这种个性信息命名为“声纹”，但依然在文章的最后提出了这个问题：成年人的语音是否会随着时间发生明显的变化？如果是，那么将会如何变化？

20世纪70年代以来，尤其是最近二十年，随着模式识别和机器学习理论的发展，自动说话人识别技术（Automatic Speaker Recognition）也日渐成熟。研究人员从语音中提取特征，如线性预测倒谱系数（LPCC, Linear Prediction Cepstral Coefficients）（Atal, 1976）、梅尔频率倒谱系数（MFCC, Mel-Frequency Cepstral Coefficients）（Davis and Mermelstein, 1980; Furui, 1981; 甄斌等, 2001; Kim and Sikora, 2004; Sinha *et al*., 2005）、感知线性预测（PLP, Perceptual Linear Prediction）（Hermansky, 1990; Tranter *et al*., 2004）等，并通过一定的建模方法，如高斯混合模型-通用背景模型（GMM-UBM, Gaussian Mixture Model-Universal Background Model）（Reynolds, 2000）、高斯混合模型-支持向量机（SVM, Support Vector Machine）（Wan and Campbell, 2000; Kharroubi *et al*., 2001; Campbell *et al*., 2006）、联合因子分析（JFA, Joint Factor Analysis）（Kenny, 2005; Kenny *et al*., 2005, 2007, 2008; Vogt and Sridharan, 2006; Yin *et al*., 2007; Liang *et al*., 2012）以及更一步的i-vector方法（Dehak and Kenny, 2009; Kenny, 2010; Senoussaoui *et al*., 2010, 2011; Dehak *et al*., 2011; Cumani *et al*., 2011; Glembek *et al*., 2011, Li *et al*., 2011）等，以得到声纹模型。

日本语音界先驱Sadaoki Furui教授总结了自动的说话人识别技术几十年来的进步（Furui, 1997），同时也指出了其中悬而未决的若干课题。而其中之一就是，如何处理语音中的长时变化。研究人员怀疑由语音中所提取的声纹是否存在系统的长时变化；如若果真存在，那么这种系统的变化将可以有助于更新声纹模型以反映声纹的渐变。

Bonastre教授等几位语音界前辈也专门发表论文（Bonastre *et al*., 2003），提出了相似的观点。他们认为唯一地刻画一个人的声音，其挑战在于声音随时间变化。这种变化或者是短期变化（一天内不同时段的变化）、中期变化（一年内的变化）或者是长期变化（随着年龄增长而带来的变化）。这种不确定的变化使得我们无法从一个人的语音中绝对断定其身份，因而他们倡导对于说话人识别技术的实际应用人们应持一个谨慎的态度。

而在实际的说话人识别系统中，伴随声纹预留与声纹验证之间的时间间隔而来的系统性能下降，在文献中也屡次被提及。

Frank Soong等使用两个月内的五次录音，在100人（男女各半）规模的语音库上进行了若干组实验（Soong *et al*., 1985），得出了结论：用于测试的语音样本与训练声纹模型所使用的语音样本之间时间间隔越久，说话人识别系统的性能越差。他们需要不断的更新矢量量化VQ码本来维持系统的性能。

日本的两位学者Kato和Shimizu也提到了同样的问题（Kato and Shimizu, 2003），间隔三个月之后说话人识别系统的正确率有较为明显的下降，而另一位研究人员Hebert认为“老化（Ageing）”是其中之因（Hebert, 2008）。

浙江大学CCNT实验室设计的一个声纹打卡系统（单振宇等, 2005），记录了实验室成员每天的声纹打卡情况。其中2004年3月至2005年5月15日的记录中，开始的50天里识别准备率为69.02%，而在使用间隔稍近的语音更新过声纹模型后，识别准备率提高到了74.19%。

综上，说话人识别中存在明显的时变现象，这是研究人员普遍认可的观点；同时，许多研究机构在如何应对随时变所带来的系统性能下降方面进行了各种尝试。详见下节。

1.3　时变问题的研究现状

1.3.1　研究现状概述

以一个典型的说话人确认系统为例，其框架如图1.1所示。可见系统框架中的四个重要组成部分分别为：训练数据的准备、特征的提取、模型的建立以及分数域的决策。因此，本节关于说话人识别中时变问题的国内外研究现状的综述，也是依次从这四个方面来分别展开。

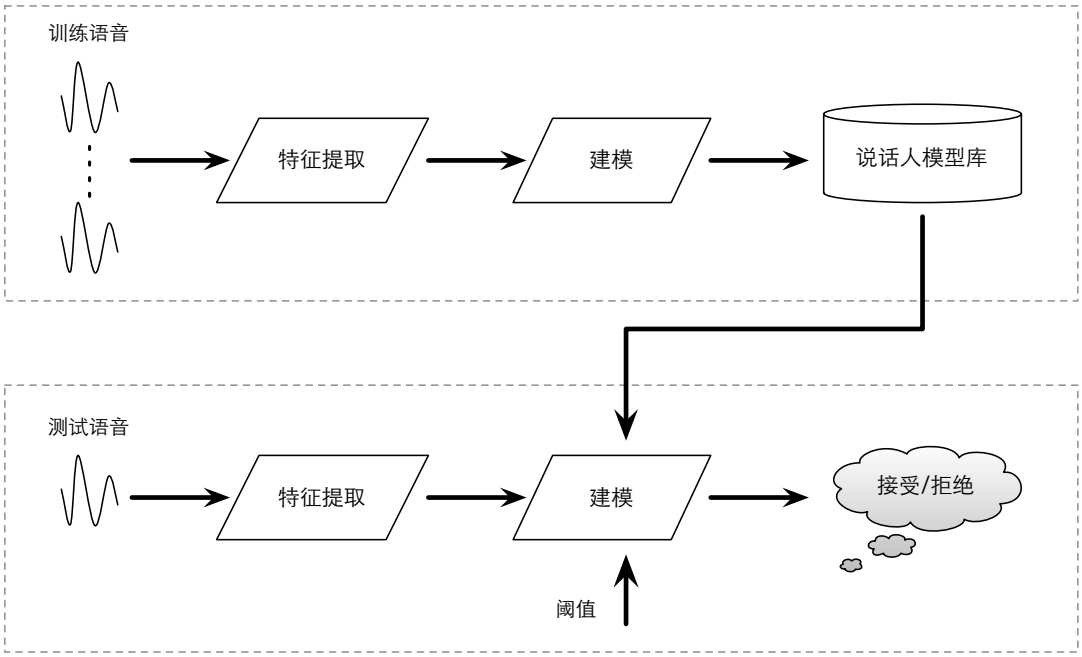

图1.1　一个典型的说话人确认系统框架

1.3.1.1　更多样的训练语音数据

采用更多样的训练数据，即所谓的“结构化（Structural）”训练方法，是缓解几乎所有模式识别问题中存在的训练与识别数据失配状况的一种最常见的做法。从机器学习的观点来看，更多样的训练数据就意味着更具有代表性的模型。

因此，一些研究人员选用一段时间内的若干次会话语音作为训练数据，以应对语音的长时变化（Soong *et al*., 1985; Bimbot *et al*., 2004）。

Markel和Davis在他们早期进行自动的说话人识别研究时，曾在一篇论文中使用了依次间隔至少为一周的连续五次会话的语音来作为参考集（Reference Set, 即训练数据）（Markel and Davis, 1979），获得了最好的识别结果。这种做法固然简单方便，其缺点也是显而易见的：建立说话人模型之前会经历相当一段时间的语音采集阶段。对于一个实际的系统而言，需要一个较长的用户注册过程，用户接受度也会降低。

Beigi使用了更易于接受的一种方案，称之为“数据增强（Data Augumentation）”方法（Beigi, 2009, 2010）：用户注册是一次完成的，但在之后系统使用的过程中，每次系统判定为“接受”的语音都会追加到原始的训练语音数据之上，从而训练得到一个更加“与时俱进”的说话人模型。论文中的实验结果表明，这种方法在最初较短时间的测试中比通常不追加数据的方法要稍差（错误率约增加5%），但在随后的测试中，随着时间间隔的增加，数据增加方法的优势逐渐明显（错误率由43%下降到了32%）。由于追加操作对用户不可见，故不会影响用户的接受度。

但这种方法必须保留原始的训练语音和系统判定为“接受”的语音数据，以完成随后所有可能的追加训练语音操作来更新模型。对于系统而言，这是一笔额外的存储成本。同时，系统判定为“接受”的语音，并不一定就是目标说话人的真实语音，于是“接受”阈值的选取就变得极为关键。

1.3.1.2　更稳定的特征

浙江大学CCNT实验室的研究人员对声纹打卡系统几年来采集到的语音数据进行了初步的探索（陈文翔等, 2010）。他们利用PRAAT工具（Boersma, 2002）分析了语音的基音频率（Pitch、Fundamental Frequency）、能量以及共振峰（Formant）的变化，得到了不同的语音参数随时间变化的分布情况。其实验结果表明，基音频率范围会随着时间推移而出现明显的上下波动，而能量的均值却趋于稳定，较少受时变影响。他们认为系统识别率的下降可能与共振峰的改变有关系，因为共振峰与系统所采用的MFCC特征直接相关联。

中国科技大学陆伟博士的研究也表明基音频率在不同时间的随机变化是一个重要的因素（陆伟, 2008）。对于高基音频率的语音，MFCC特征受其影响，可能会降低系统的时变鲁棒性。基于此陆伟博士提出一种称之为SMFCC的新特征。这种特征的提取过程如下：首先对语音的幅度谱进行平滑操作，并求取其谱包络，然后经过梅尔（Mel）滤波再求其倒谱系数，以尽量降低基音频率之影响，使得参数可以更加准确地刻画声道的特性。SMFCC特征在说话人识别系统中的表现与性别相关。对女性说话人尤其有效，其误识率远远小于MFCC特征的系统；然而对于男性说话人，由于其基音频率较低，其MFCC特征受影响的程度远低于女性说话人，所以两种特征性能相当。

1.3.1.3　更有代表性的自适应说话人模型

既要逐步更新说话人模型以使其保持代表性，同时又要避免额外的语音数据存储开销，很自然地，研究人员转向寻求合适的说话人模型自适应方法。模型自适应是指已有模型的参数同测试语音分布之间存在失配时，通过调整模型分布使模型参数和语音尽可能相匹配的一类方法。最大后验概率（MAP, Maximum *A Posteriori*）（Lee *et al*., 1991; Lee and Gauvin, 1993; Gauvin and Lee, 1994; 李虎生等, 2003）和最大似然线性回归（MLLR, Maximum Likelihood Linear Regression）（Leggetter, 1995; Leggetter and Woodland, 1995a, 1995b）是语音领域最常见的两种说话人自适应方法。

Beigi 采用了 MAP 方式，将判断为“接受”的语音（提取特征作为观测向量）

在原始的说话人模型（高斯混合模型）上进行 MAP 自适应，从而更新说话人模型（Beigi, 2009, 2010）。MAP 自适应是指在假设模型参数为参数空间中的随机向量的基础上，通过利用贝叶斯（Bayesian）定理对模型参数引入先验知识，从而减少可靠估计模型参数所需要的数据量的方法。 在说话人识别领域，通常使用 MAP 方法对 GMM 模型中每个高斯混合成分的均值进行自适应，更新方式如下：

$$\hat{\mu} = \frac{\tau\mu_0 + \sum_{t=1}^{T}\gamma(t)o_t}{\tau + \sum_{t=1}^{T}\gamma(t)}. \qquad (1\text{-}1)$$

其中，$\hat{\mu}$ 指的是经过MAP算法更新后的模型参数，μ_0 是原始模型参数（先验信息），T 是新语音的帧数，o_t 为新语音的第 t 帧特征，$\gamma(t)$ 为此高斯在 o_t 上的加权概率（所有高斯混合概率密度函数进行加权），τ 是一个元参数，它控制了自适应对先验信息的依赖程度。τ 取值越大自适应后的说话人模型参数越接近于原始模型参数，τ 取值越小自适应后的说话人模型参数则更接近新语音的各帧特征分布。通常会选取2到20这个范围内的数值。

Beigi的实验表明，经过MAP自适应后的系统性能有了大幅提高，相隔三个多月的测试结果显示，错误率由自适应之前的43%下降到了18%。

Lamel 和 Gauvin 采用了另一种 MLLR 的自适应方法（Lamel and Gauvin, 2000）。MAP 算法中，只对有相应自适应数据的参数分布进行调整，所以当自适应数据量较少时效果有限。而 MLLR 方法从这一问题出发，利用最大似然准则对高斯混合成分的参数估计线性回归变换，并通过对不同参数的变换矩阵进行不同层次的共享来实现在有限数据中对所有模型参数进行自适应的目的。对每组共享的参数，MLLR 只估计一个变换矩阵，对该组中所有参数进行统一的线性变换以实现对模型参数的更新。MLLR 方法对高斯混合成分的均值的更新方式如下：

$$\hat{\mu} = \mathrm{A}\mu + b. \qquad (1\text{-}2)$$

其中，$\hat{\mu}$ 指的是经过 MLLR 方法更新后的模型参数，μ 是原始模型参数（先验信息），A 是一个 $n \times n$ 的线性变换矩阵（n 是特征的维数），b 是一个 n 维向量。这个等式一般也会写成如下形式：

$$\hat{\mu} = W\xi. \qquad (1\text{-}3)$$

其中 W 是一个 $n \times (n+1)$ 的矩阵，而 ξ 是扩展的原始模型参数向量，定义如下：

$$\begin{cases} \mu^T = \begin{bmatrix} \mu_1 & \cdots & \mu_n \end{bmatrix} \\ \xi^T = \begin{bmatrix} 1 & \mu_1 & \cdots & \mu_n \end{bmatrix} \end{cases}. \tag{1-4}$$

MLLR方法的核心就在于W矩阵的估计，通常采用EM算法来估计，使得新语音数据（各帧特征）的似然最大化。

Lamel和Gauvin的实验结果表明，通过使用干预语音（即前述新语音）来进行MLLR自适应，说话人确认系统在最后两次会话（Session）的等错误率（EER, Equal Error Rate）由之前的2.5%下降到了1.7%。

两种模型自适应方法在各自所用数据库上均取得了很好的效果，并且除了每次自适应操作的运行开销，不需要额外的语音存储开销。但与之前相似，自适应的方法依然面临着如何选取一个合适的“接受”阈值的问题。若阈值选取过低，系统的错误接受率（False Acceptance Rate, FAR）将会提高，更多的假冒者语音被用来更新目标说话人模型，系统的安全性会大大降低；而若阈值选取过高，系统的错误拒绝率（False Rejection Rate, FRR）将会提高，目标说话人模型的更新频率降低，无法起到应对时变影响的作用，用户体验无法改善。

1.3.1.4　更合适的分数域决策

Kelly等研究人员以英国广播公司（BBC）针对18位普通人跨度长达60年的访谈纪录片作为时变数据，进行了分数域的研究（Kelly and Harte, 2011; Kelly *et al*. 2012, 2013）。他们发现，目标说话人的真实语音（称之为True Speaker），随着目标说话人年龄的增长，在目标说话人模型上的打分有着明显的下降趋势；而假冒者的语音（称之为Impostor），随着假冒说话人年龄的增长，在目标说话人模型上的打分则变化不明显，较少被时间因素影响。于是随着年龄增长，真实语音得分与假冒语音得分差距越来越小，系统的等错误率就随之增大。

基于上述观察，Kelly等提出了一种与时间相关的决策边界算法。采用一种称之堆叠分类的方法，将待识别语音在目标说话人模型上的似然分与时间间隔组成“对（Pair）”，利用SVM依据大边距准则找到在结构风险最小化意义下最优的分类边界对“对”进行二次分类。

实验结果表明，当跨度达到60年时，通常的固定阈值决策方法下系统的错误率会由最初的10.8%上升至36.1%，而采用了时间相关的决策边界算法后，系统的错误率由最初的7.3%仅上升至17.5%。说话人识别系统的时变恶化程度得到了极大的缓解。但这种方法本质上相当于为似然分的每个维度分别估计了一条随时间变

化置信度阈值下降的曲线，随着时间的推移错误接受的情况会显著增加；而对传统固定阈值算法，随时间变化系统不再在原有的等错误率点，主要增加了错误拒绝。故采用堆叠分类算法后，系统的错误接受增加，安全性下降。

1.3.2　研究现状分析

几十年来，说话人识别研究中的热点问题一直集中于跨信道（声纹预留与声纹验证时信道不一致）、背景噪声减弱或消除、多说话人（一段语音中含有多个不同说话人）分割、短语音（特殊应用条件下，用于提取声纹的有效语音时长过短，甚至少于2秒的情况）识别以及多发音方式（声纹预留与声纹验证时说话人发音方式不一致，如情绪、语速、音量、语种等）混合等问题，并且说话人识别领域的国际评测，如美国国家标准与技术研究院（NIST, National Institute of Standards and Technology）组织的说话人识别评测（SRE, NIST Speaker Recognition Evaluation）（NIST, 1995），多年来关注的重点也一直是在跨信道和多说话人等方面。直到最近几年来，随着说话人识别技术在实际系统中越来越多的应用，时变问题才逐渐被越来越多的研究人员所关注。如前文所述，许多研究机构在时变方面已有不少尝试，但仍然存在一些问题。

与其他很多的训练和识别失配的问题相似，模型更新依然是应对时变问题的一种最简单且有效的方式。对于一个实际的说话人识别系统，最理想的情况是，用户每隔一定或不定的时间间隔，登录进入系统来更新自己的声纹模型。这样既保证了每次模型更新使用的都是身份无误的语音，且模型能够保持“与时俱进”。但实际情况是，在很多应用中并不能随时得到说话人当下的身份无误的语音。同时，反复地登录进系统更新自己的声纹模型会给用户造成太大的额外负担，而这种额外负担对于说话人识别系统本身的易用性和推广度会造成致命的伤害。

基于这种考量，前文中所提到的模型更新方法，无论是通过更多样的训练数据直接进行“结构化”模型训练，还是利用相隔更近的时间点所得到的语音数据在原始说话人模型上进行自适应，都可回避对于用户增加的额外负担。当然在回避用户额外负担的同时，不可避免地，这些方法都对说话人识别系统提出了更高的要求，要么为了积累语音数据样本从而需要一个较长的用户注册过程，要么虽然用户注册过程一次完成但需要系统能够“相对精确”地选取更新操作的参数——“阈值”。因此说话人识别系统，要么由于注册过程长带来的繁琐或者由于“阈值”偏高带来的更新频率较低，而造成用户体验下降；要么由于“阈值”偏低带来的可能错误误更新，而造成安全性下降、潜在的风险提高。

此外，这些模型更新的方法，撇开系统安全性方面的考量，从某种意义上来

讲，更像是一种“盲目”的更新策略。它们并没有触及到时变现象的本质，而只是“盲目”地通过各种方式向声纹模型中追加时间间隔更近的语音数据，以期缓解训练与识别的失配状况。与之相反，SMFCC特征提取方法和时间相关的决策边界算法，分别考虑了时变问题对于基音频率和分数域的影响及趋势，给出了极有针对性的解决策略。虽然二者各有局限，比如前者对基音频率的高低比较敏感、因而对于男性说话人效果并不明显，后者也只是从分数域的变化趋势上解释了系统为何随时间变化而性能恶化这一现象、并没有从根本上回答声纹时变的问题，但是这种“针对性”的尝试理应成为时变课题研究的方向。

勿庸置疑，任何“针对性”策略的得出都是建立在前期大量数据分析的基础之上，而目前研究中存在的一个关键问题是没有一个“公认”的“专门”用于说话人识别中时变课题研究的时间跨度较长且人数尚可的声纹数据库。研究人员分别使用各自由于种种客观原因而准备的声纹库，这些声纹库或者时间点采集较少、时间跨度较短，或者人数有限（几人或十几人规模），或者录制环境变化明显（信道、说话方式等）等。这些变化因素给时变研究带来了不同程度的困难，使得算法推广性较差。

1.3.3　时变问题研究难点

从研究现状的分析可以看出，说话人识别中的时变问题，其研究难点主要有以下两个方面。

1.3.3.1　专门的时变声纹库

数据采集是解决一切模式识别问题的基础，时变说话人识别研究亦不例外。语音的录制一般会受到各种外界环境、说话人自身因素以及说话内容的影响，因此一个排除这些种种干扰、专门用于时变研究的声纹库的录制就显得尤为关键和必要。这种专门的声纹库对于时间跨度、录制会话次数以及参与人员数目等方面也都应有相当的要求，因此相对于普通的声纹库，其录制难度大大增加。某种程度上，这也是截止目前时变问题相关研究较少的一个重要原因。

1.3.3.2　针对性的声纹时变规律

探索声纹的时变规律，并依此规律改善说话人识别系统的时变鲁棒性，是时变问题研究的核心和本质。与“盲目”地追加用户语音数据以更新其声纹模型相反，知晓了声纹的时变规律，就可以“有指导” 、“针对性”地设计出更加稳定的声纹特征，进一步训练得到时变表现更加稳定的声纹模型；在现有的经典的说

话人识别框架下，最终提高整个说话人识别系统的时变鲁棒性。这里需要特别说明的是，声纹的时变规律，并不等同于单纯语音的时变规律，因此规律的探索应是在说话人识别的基本框架之下。

1.4　研究工作概述

1.4.1　研究思路

本文的研究目标是在不影响整体识别性能的前提下，提高现有说话人识别系统对于时间变化的鲁棒性。而针对说话人识别中时变问题研究的两个难点，本文的研究思路是：以一个合适的长期专门录制的时变声纹库作为基础，在现有的经典的说话人识别框架之下，探索声纹的时变规律，研究对于时间变化鲁棒性较好的声学特征表达方式及其与声纹模型相结合的优化方案。整体研究思路如图1.2所示。

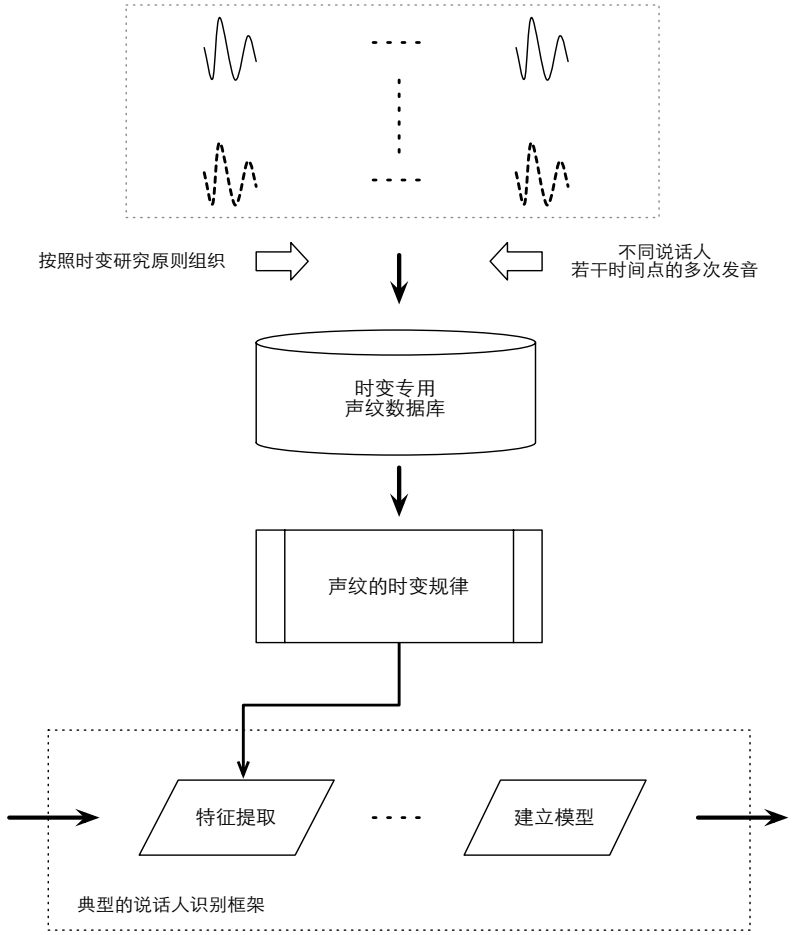

图1.2　论文研究思路示意图

本文对于说话人识别中时变问题的研究被分解为三个子问题，即如何构建一个合适的长期的时变专用声纹数据库、从哪个合适的角度来探索声纹的时变规律、以及如何把该规律应用于典型的说话人识别框架之中。关于这三个子问题的研究思路展开如下。

1.4.1.1　构建合适的时变声纹库

何为一个合适的时变声纹库？既然研究的目的专注于说话人识别中的时变问题，那么很显然，理想情况下在构建声纹库时，时间应该是该数据库中唯一变化的因素，除此之外的其他任何因素均应保持不变。现实情况下很难做到这一点，毕竟“人不可能两次踏入同一条河流”（赫拉克利特语），因此很难保证说话人的两次录制会话之间外部环境和人自身的各种条件因素完全一致。

在这种情况下，抛开不可控因素，我们应当保证所有其他可控的外界因素尽量保持一致，例如，所有录制会话在同一地点同一录音设备同一录制信道下完成、背景噪声尽量保持在一个恒定的状态。

除了这些可控的外界因素外，人自身因素也是一个不可忽视的重要方面。比如如何保证说话人在各次录制会话间的心理和生理状态基本维持在一个没有明显差异的水平等，这种要求更高。除了依赖说话人自身的把握以及录音负责人员的合理组织和安排外，还可以从录制会话的形式和内容方面做些工作。比如自由交谈这种录制形式，语音内容不固定，而且自由交谈的方式下说话人的情绪等方面会容易受到会话搭档的影响，这样对于时变声纹库就不是一种合适的录制形式。相反朗读这种录制形式，说话人可以有效控制自己不同录制会话间的发音方式，基本维持在正常的朗读状态下，对于时变声纹库就是一种合理的录制形式。朗读文本的设计尽量选择情感倾向性不大的新闻文本，固定文本内容。这样也会大大减少同一个说话人各次不同录制会话间可能存在的差异。

总之，时变声纹库构建的思路就是除了时间这个目标变化因素外尽最大可能保持其他各因素的稳定。

1.4.1.2　探索声纹的时变规律

声纹被认为是语音中所包含的说话人个性信息，因此声纹随时间的变化与人的语音随时间的变化密切相关。

生理解剖学上曾有研究人员称语音为“人类年龄的镜子（a mirror of age）”（Segre, 1971; Orlikoff, 1990），正说明了语音的时变特性（vocal aging）（Ptacek and Sander, 1966; Shipp and Hollien, 1969; Ryan and Burk, 1974; Hartman, 1979; Hartman

and Danhauer, 1979; Horri and Ryan, 1981; Linville and Fisher, 1985; Mueller, 1989）。通常将年龄段大致分为四个时期：儿童期、青春期、中青年期和老年期（侯丽珍, 2010），各个年龄段的特点如下：

儿童期指的是从出生至青春期前，整个发声系统（包括呼吸器官、声带、共鸣器官和构语器官等）尚处于发育过程，语音的基音频率（F0）较高，稳定性较差，基音频率微扰（jitter）和振幅微扰（shimmer）都较大，同时共振峰频率（F1、F2、F3等）也较高，其中新生儿基音频率最高。10岁以后逐渐进入青春期，青春期最显著的变化在于变声期的到来。变声初期语音基音频率不稳，变化范围较大，微扰一般也较大；但随后语音的基音频率降低明显，语音渐趋稳定，基本变为成人语音。而在中青年阶段，语音的基音频率随年龄增长会稍有下降，基音频率的范围变小，微扰则都有所增大，但还在正常的范围内。而除此之外，研究表明多数其他语音参数并没有特别显著的差异，语音处于大致比较稳定的状态。而进入老年后，发声系统各个器官结构逐渐老化，同时功能衰退。研究认为，男性说话人的语音基音频率，尤其是60岁之后一直呈上升的趋势；而女性说话人除了在绝经期由于性激素的变化语音基音频率上升显著，其他阶段直到老年基音频率一直较稳，呈现下降趋势。老年语音还有稳定性差和微扰较大的特点。

由此可见，生理解剖学上的多数研究都表明了，即便是在个体语音特征表现相对最稳定的中青年时期，男性和女性说话人的基音频率都随时间变化呈现出一个缓慢下降的趋势。Reubold等人在男女成年人语音上的长期研究也表明，基音频率和第一共振峰（F1）的变化趋势和速度大致相当（Reubold *et al*., 2010）。类似的研究还有很多，这里不再一一赘述（Linville, 2004; Rhodes, 2011; Stathopoulos *et al*., 2011）。同样地，国内的研究人员，例如浙江大学CCNT实验室研究人员（陈文翔等, 2010）和中国科技大学陆伟（2008），他们的工作也证实了基音频率存在随机变化的现象。

基音频率指的是当发浊音时气流通过声门使声带发生振动而产生的准周期激励脉冲串的频率，它反映了声带的特性；而共振峰指的是，声带产生的声波在传播过程中经过由喉腔、咽、口腔和唇腔、鼻腔等共鸣器官所组成的声道，由于共振现象使得频域中不同频率的能量重新分配，而形成的能量强的部分，它反映了声道的特性。因此二者是语音中所包含的说话人个性信息（即声纹）的重要体现。

随着时间的变化，基音频率和第一共振峰所反映出来的下降趋势，从更广义的角度来讲，这种趋势可以被看作是语音中所包含的说话人个性信息在语音中不同频带上分布状况的一种变化。也就是说，我们认为，语音中各个频带，除了可能包含有语义信息、语种信息、稳定的说话人个性信息（声纹的稳定部分）等等

之外，还存在一种与时间相关的说话人个性信息（声纹的时变部分）。

因此，本文关于声纹的时变规律的探索将从频带的角度展开。Lu等人曾经从频带能量入手进行研究，发现说话人个性信息在语音中不同频带上的分布是不均匀的（Lu and Dang, 2007, 2008）。于是一种可能的假设是，这种与时间相关的说话人个性信息，在语音中不同频带上的分布也可能是不均匀的。时间相关个性信息分布较多的频带，对于说话人识别系统的时变鲁棒性自然是不利的。因此对于各个频带，我们可以在兼顾声纹的稳定部分和时变部分的前提下，确定其对说话人识别系统性能的贡献程度。

这种贡献度既可以从单纯语音特征（如频带能量分布）的角度来，又可以是性能驱动式的，以最终的说话人识别结果来反馈其贡献。因为很多情况下，特征与模型作为一个模式识别系统的两个最主要模块，二者紧密结合并相互影响：特征的作用需要通过模型来完善，模型的性能亦会受到特征的制约。对现有的模块化的模式识别系统进行特征模块与模型模块的联合优化可以有效提高系统性能。

1.4.1.3　设计鲁棒的说话人识别系统

作为一个典型的模式识别任务，说话人识别最核心的问题依然是特征（Huang *et al*., 2001）。而以MFCC为代表的倒谱系数依然是说话人识别中广泛应用的一类特征，其提取过程大致如图1.3所示。

前文中也有提到，并不是语音中包含的所有信息在区分说话人身份上都有贡献。因此对于说话人识别系统来说，一种理想的特征应该是“在不同说话人之间具有比较大的差异，但说话人自身来讲差异较小，……，并且不受语音中的长时变化所影响，……”（Wolf, 1972; Rose, 2002; Kinnunen and Li, 2010）。而对时变说话人识别研究而言，其目标就是从说话人的语音提取出说话人个性信息相关，同时对于时间变化并不敏感的信息来作为特征。于是从频带出发，在特征提取的过程中，应该强调那些对说话人个性信息表现出较高区分度、同时对时间相关信息表现出较低区分度的频带，而弱化表现反之的频带。

在上述倒谱特征中，与频带的强调和弱化直接相关的核心在于三角滤波器组一前一后的两处设置——滤波前的频率弯折和滤波后的输出加权。

图1.3中的三角滤波器组是均匀公布的，与弯折后的频带相对应；MFCC的频率弯折方式如图1.4所示。可见梅尔刻度（Mel-scale）变换的MFCC加强了较低的频带，即低频分辨率较高，而弱化了较高的频带，即高频分辨率较低。因此通过合适的频率弯折方法，可以在一定程度上达到强调或者弱化指定频带的作用。

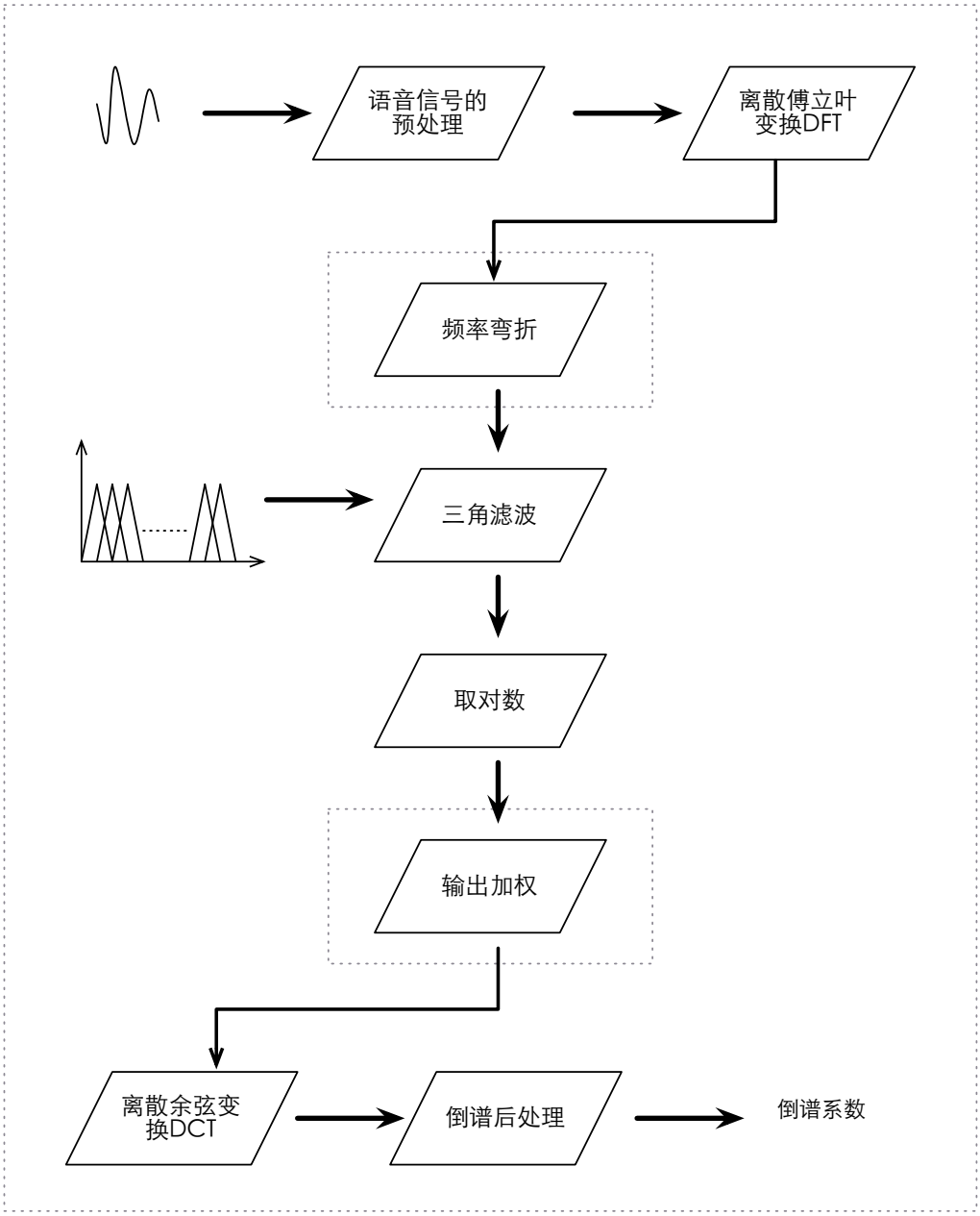

图1.3　倒谱系数计算流程示意图

类似地，在滤波之后、离散余弦变换（DCT, Discrete Cosine Transform）之前，对各个三角滤波器的输出（对数能量）进行加权，也是一种很直接地，在最终生成的倒谱系数中强调或者弱化相应频带影响的方式。

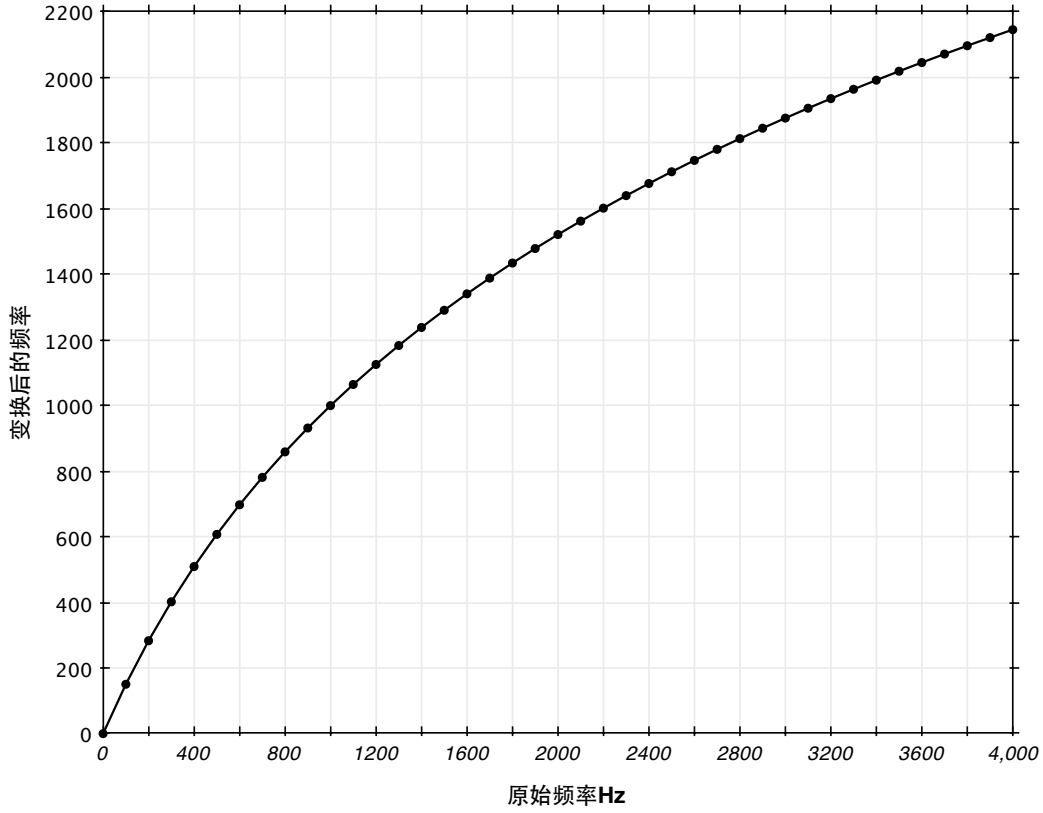

图1.4　MFCC的频率弯折示意图

1.4.2　工作内容

本论文以经典的基于倒谱系数特征和GMM-UBM建模方法的说话人识别框架为基础，工作内容包括以下几个方面，如图1.5所示。

1.4.2.1　时变声纹库

为了开展时变说话人识别研究，本文建立了一个60人规模、时间跨度长达三年、预计录制会话16次的时变声纹数据库Chronos。

所有语音数据均在实验室内部的一个专门录音间完成，录音设备、信道及录音软件均保持不变，录音环境背景噪声基本维持在一个恒定的水平，录音过程中无突发噪声影响。

录音文本来自于人民日报中的新闻语料，由100条汉语句子构成，每句话包含字数不定，一般为8到30字之间。录音过程中要求录音人员以朗读的方式逐句读出录音文本，录音文本在所有16次录制会话期间保持不变。在各次录制会话间，采用了梯度间隔的做法。最初的录制会话间隔较短，一周或者一个月等，之后录制会话间隔逐渐增加，两个月、四个月、半年等。

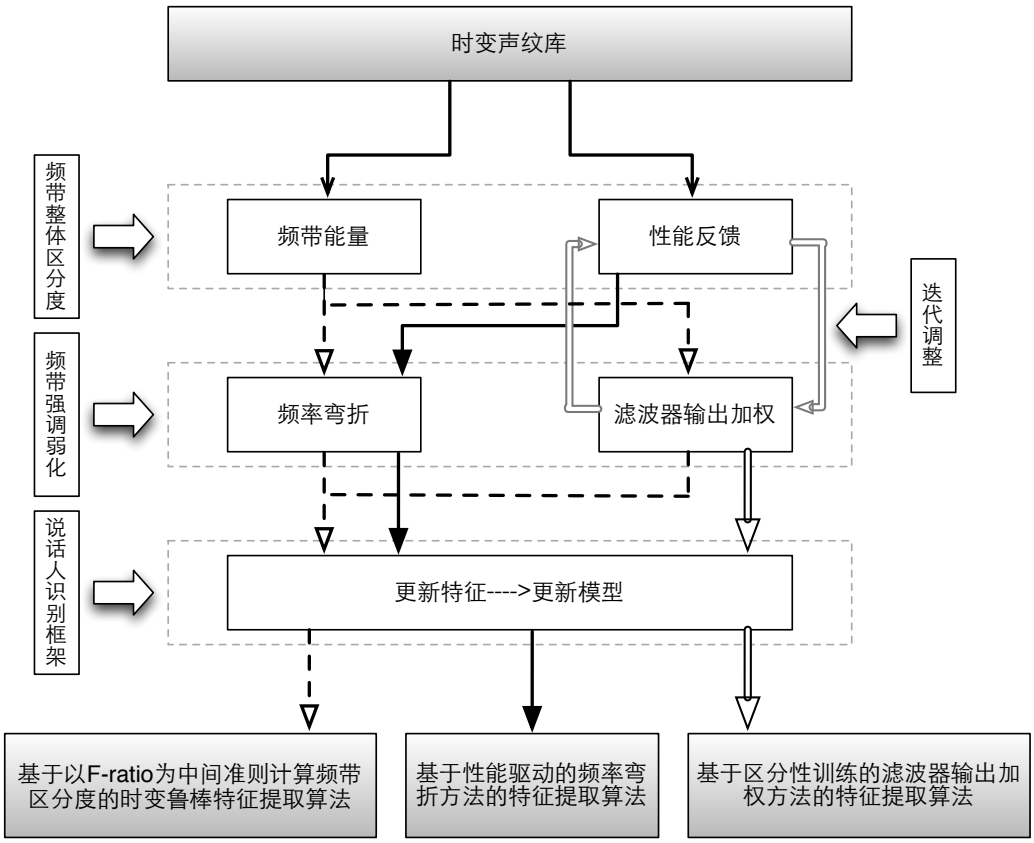

图1.5　论文工作内容示意图

考虑到长期录音的可行性，录音人员为从校园里招募的普通话流利的学生：30名男生、30名女生，年龄在20岁左右，来自于不同的院系。录音时间点的设置也充分考虑到了假期安排等因素。

1.4.2.2　基于以F-ratio为中间准则计算频带区分度的时变鲁棒特征提取算法

这里的频带区分度指的是既考虑频带对于说话人个性信息的区分度，同时也考虑其对于时间相关信息的区分度。综合考虑二者关系，就可以得出每一频带对于时变说话人识别这一任务的整体区分度。这里的时变鲁棒性算法指的是1.4.1.3节所提到的两个方面：特征提取过程中的频率弯折和滤波器输出加权。易知，区分度高的频带，所对应频率弯折的分辨率就应该越高，类似于图1.4所示，或者对应滤波器的输出应给予的权重应该越高。

如何得到整体区分度曲线？本算法中尝试了基于频带能量的F-ratio准则。

Lu等人的工作曾利用F-ratio准则来计算不同频带对于说话人个性信息的区分度（Lu and Dang, 2007, 2008）。于是更改下数据组织模式，同样可以利用F-ratio作为中间准则来计算不同频带对于时间相关信息的区分度。因此以频带能量作为

参数，对于每个频带，存在两种F-ratio值，说话人个性信息F-ratio和时间相关信息F-ratio。显然对于时变说话人识别这个任务来说，两值相比会是一个比较合理的整体区分度指标。这样整体区分度曲线就可以得到了。

1.4.2.3　基于性能驱动的频率弯折方法的特征提取算法

在以F-ratio为中间准则计算频带区分度的时变鲁棒性算法中，利用了频带能量进行区分度计算，但在实际系统中频带能量的区分度还会受到模型设计、参数优化等因素的影响，因而基于频带能量的F-ratio准则区分度高与系统实际性能好是否总是一致也是一个问题（将在后文进行相关研究）。因此从性能驱动的角度出发，利用实际得到的系统性能来评估频带区分度则更为直接。

本文详细探讨了时变说话人识别系统性能的综合评价指标，并在此基础上，针对频率弯折方式的特点，设计了性能驱动的准则，具体作法是：对于某一指定的频带，保持其他所有频带分辨率不变，唯独加强该频带，设计出一个说话人识别系统，该系统的性能就作为该指定频带的整体区分度指标。如此经过一系列的说话人识别实验，就可以得到整体区分度曲线，并在其基础上进行频率弯折，从而得到更为鲁棒的系统。

1.4.2.4　基于区分性训练的滤波器输出加权方法的特征提取算法

在前一小节所述的性能驱动准则的方法中，需要依据识别结果加强某一特定频带以找到对说话人具有良好区分性但对时变因素较不敏感的频带，并对其利用人工选定的参数进行加强，这样的方法还存在一些可以改进的方面：首先，频带加强、弱化的程度受人工干预；其次，仅利用识别结果对模型参数进行了一次评估（相当于在图1.5中右侧只有一次性能反馈），没有充分体现系统特征模块和模型模块间的相互影响。以上两点使得性能驱动准则方法很难保证找到可以使系统性能达到局部最优的特征参数。第三，该方法需要遍历某一或某些频带的所有组合并重复进行试验，这样会耗费大量的计算资源。

为解决上述问题，提出了使用区分性特征提取算法来进行时变问题局部最优特征参数的方法。算法起初将各滤波器输出权重设定为某一初始序列（一般为等权重），经过说话人识别系统的建模和打分过程，得到系统初始的错误率；而后以最小分类错误（MCE, Minimum Classification Error）为准则，调整各滤波器输出权重，重复上述建模和打分过程；迭代若干次，直到找到一个性能比较好的权重序列。权重的大小就体现出了相应频带对于时变说话人识别这个任务的区分度，反馈过程如图1.5所示。

为了使算法更适合寻找适合时变问题的特征参数，提出了最小化会话方差（MSV, Minimum Session Variance）的准则。MSV准则利用与MCE类似的技术将分类错误嵌入光滑可导的Sigmoid函数计算错误率，不同之处在于最终的目标函数不再是所有错误率的简单加和而是所有会话各自错误率间的方差。这样，通过寻找使得MSV最小化的特征或模型参数，就可以使系统对时变现象更加鲁棒。因此，将MSV与MCE准则组合使用，找出既使得说话人个性信息较为突出，又对时变现象较不敏感的频带。

1.5　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。首先介绍了说话人识别技术及应用背景，引出技术走向实用无法回避的时变问题，接着综述了说话人识别中的时变现象及国内外研究现状，然后分析了时变研究目前存在的问题及研究难点，最后阐述了本文的研究思路和大致的工作内容。

第 2 章是时变声纹数据库 Chronos 部分。首先介绍了现有的时变声纹资源及存在的相关问题，接着阐述了本文所构建 Chronos 的原则，然后详细说明了数据库的各项具体录制方案，最后从具象的频谱特征、声纹特征相关度以及系统识别率的变化等方面展现出了 Chronos 上的时变现象。

第 3 章从频带区分度谈起，提出了在时变说话人识别的任务中频带整体区分度的概念，并简要阐述了区分度的确定准则。着重从基于频带能量和 F-ratio 的准则入手，详细介绍了频带整体区分度的计算。确定了频带的整体区分度之后，就要在特征提取时强调那些对说话人个性信息区分度高且对时间相关信息区分度低的频带，而弱化表现反之的频带。本章探讨了与滤波器组设置相关的两种强调和弱化的鲁棒性算法：频率弯折和滤波器输出加权。

第 4 章从基于频带能量和 F-ratio 准则的局限性出发，提出利用性能驱动的方式来确定频带整体区分度的可行性，并且探讨了针对频率弯折方式的性能驱动准则：单独加强某一频带构建说话人识别系统，并将系统性能作为该频带的整体区分度。之后针对时变说话人识别任务的特点，定义了系统的性能指标参数。据此确定了性能驱动准则下频带的整体区分度。

第 5 章是时变说话人识别的区分性特征提取算法，它是滤波器输出加权方式下的性能驱动准则的应用。首先提出了利用区分性训练的思想进行时变说话人识别特征参数优化的思想，接着探讨了参数优化的具体准则及训练算法，同时阐述

了特征与模型参数进行联合优化的方案，最后依据说话人识别模型的结构特点进行了加速处理。

第 6 章是论文工作总结及展望部分。

第2章　时变声纹数据库Chronos

2.1　引论

2.1.1　语音资源联盟概述

著名理论物理学家詹姆斯•比约肯（James Bjorken）教授 2004 年荣获国际理论物理中心（CITP）颁发的理论和数学物理领域最高荣誉狄拉克奖章时，在颁奖典礼上做了题为“Data Matters（数据关系重大）”的报告，特别强调了实验数据对高能物理学发展所起的作用（周顺忠等, 2006）。理论物理学尚且如此，就更不必说语音研究这门典型的实验科学了。语音数据是语音技术研究的关键，这是研究人员的共识，很多语音研究机构也一直致力于语音资源的建设和搜集。但语音的录制是一件硬件要求较高，同时又非常耗费时间和资金的事情。因此，资源联盟这种组织形式就应运而生，便于联盟内部成员共享彼此出于不同目的建立和搜集的语音数据。以下是全球不同地区几个主要的资源联盟介绍。

语言资源联盟 （LDC, Linguistic Data Consortium）（LDC, 1992）是语音领域最有影响的一个开放资源联盟。在美国国防部先进研究项目局 （DARPA, Defense Advanced Research Projects Agency）和美国自然科学基金 （NSF, National Science Foundation）信息与智能系统 （IIS, Information & Intelligent Systems）部的支持下，LDC 于 1992 年成立，主办机构是宾夕法尼亚大学（University of Pennsylvania），其联盟成员包括 100 多家大学、公司以及政府科研机构等。目前语音相关的数据库约 260 多个，可为语音领域绝大多数研究提供数据支持，约有 197 个成员机构和 458 个非成员机构使用 LDC 数据进行研究。

1995 年欧洲也成立了类似的资源联盟。欧洲共同体委员会董事会第十三分组 DG XIII，特别是其语言工程部门，推动成立了欧洲语言资源联盟（ELRA, European Language Resources Association）（ELRA, 2008）。ELRA 根据卢森堡大公国法律注册成立，但总部设在法国巴黎。其成员包括大公司及主要的研发实验室，主要面向欧洲进行大规模的资源建设与共享。现有语音相关的数据库约 500 个。

日本的资源联盟建设始于 1999 年，称之为言语资源协会 （GSK, Gengo-Shigen-Kyokai，日语直译为 Language Resource Association）（GSK, 2006），但发展并不顺利，2003 年转为非政府组织 NGO，之后发展重心逐渐转向单纯文本资源方面。2006 年日本国立情报学研究所 （NII, National Institute of Informatics）发起成立了语音资源联盟 （SRC, Speech Resources Consortium）（NII, 2007b）。

与 LDC 和 ELRA 同时包含文本和语音两类资源不同，NII-SRC 是专门针对语音的一个资源联盟，至今已有 50 个左右的数据库。NII-SRC 面向亚洲范围，目前以日语资源为主。

中文语音方面较有影响的两个资源联盟分别是中文语言资源联盟 （CLDC, Chinese Linguistic Data Consortium）（CLDC, 2004）和国际中文语言资源联盟（CCC, Chinese Corpus Consortium）（CCC, 2004）。

CLDC 是由中国中文信息学会语言资源建设和管理工作委员会于 2003 年发起，由该领域的科技工作者自愿组成的学术性、公益性、非盈利性的学术团体。其中现有近 50 个语音相关数据库。

CCC 是由北京得意音通技术有限责任公司联合清华大学智能技术与系统国家重点实验室（语音技术中心和人机交互与多媒体实验室）、中国社会科学院语言研究所、香港中文大学、新加坡中文和东方语言处理学会、日本 ATR 音声言语通信研究所以及美国约翰•霍普金斯大学语言和语音处理中心等国内外重要的语音与语言研发机构发起并在 2004 年 3 月创立的。现有几十个语音相关数据库，尤其在说话人识别方面拥有十几个大规模的声纹数据库，可为说话人识别方面跨信道、发音方式变异等研究提供可靠的数据支持。

2.1.2　现有时变声纹资源

LDC 作为所有语音资源联盟中影响最大、成员范围最广的一个，近些年在时变声纹资源方面也有了相当的积累。当然受种种因素影响，更多现有（公开发表）的时变说话人识别研究所使用的数据库是由各个研究机构自行录制，或者从某些途径（如广播电视）中整理而来。现有（公开发表）时变声纹资源整理如下：

2.1.2.1　**CSLU Speaker Recognition Corpus**

CSLU 说话人识别数据库（Cole *et al*., 1998）是由美国俄勒冈研究院（Oregon Graduate Institute）的口语理解中心（Center for Spoken Language Understanding）自 1996 年 9 月开始录制的电话语音声纹库。在两年的时间内，每位参与录制的说话人被要求隔段时间打电话到数据采集系统，共计每人 12 次录音会话。间隔有几天和几周之分，具体说来，第一个月一周内采集两次电话语音，第二个月和第三个月不采集，第四个月采集一次电话语音，第五个月和第六个月不采集，依照同样的规律再重复三次上述采集过程，即得到总共的 12 次录音会话。参与录制的说话人遍布美国各州，并使用不同的电话。每次电话录音中，每位说话人都会被系统要求根据提示语来重复六个固定的词语四遍、重复八个固定的句子四遍、六个固

定的数字串（长度为 5），然后回答有关个人信息的问题，诸如眼睛颜色、出生月份等，此外还要就某两个感兴趣的问题分别自由谈论 20 秒钟，最后模仿系统发音跟读一个固定的句子。2006 年时 CSLU 通过 LDC 发布了该数据库的 1.1 版本（序列号为 LDC2006S26），共包含 91 位说话人的 12 次电话语音。

2.1.2.2 The Multi-Session Audio Research Project (MARP) Corpus

MARP 数据库（Lawson *et al*., 2009a, 2009b; Godin and Hansen, 2010）由美国罗马空军发展中心 （RADC, Rome Air Development Center）、空军研究实验室（Air Force Research Laboratory）以及绿洲系统公司（Oasis Systems）联合完成。在三年的时间内（2005 年 6 月——2008 年 3 月），在一个高度可控的无回声录音室，每位参与录制的说话人利用高质量麦克风设备进行语音采集，共计每人 21 次录音会话。论文中并没有明确说明各次录音间间隔是否固定。每次录音中，每位说话人首先朗读出十个固定的口语陈述句、以感叹或疑问语气重新读出上述句子、朗读出十个扩展长度的口语句子（其中五个保持不变，另外五个每次录音时随机地从固定集合中选取）、以耳语形式读出上述句子集合中的其中十句（各一半，保持不变），接着朗读每次录音时随机选取的一到两分钟的一个段落，最后是一段大约十分钟的自由交谈对话。对话的搭档在 21 次录音中保持固定，交谈双方被鼓励保持对于对话相等的参与度，话题不定。总共有 32 位说话人完成了全部 21 次语音采集。该数据库已有计划通过 LDC 发布。

2.1.2.3 The Greybeard Corpus

Greybeard（Brandschain *et al*., 2010）是直接由 LDC 官方组织专门为了时变说话人识别研究而准备的一个数据库。自 1995 年起，LDC 一直致力于为各类研究目的而进行各种语音采集。因此，当他们发起 Greybeard 项目时，一个很自然的想法是，从以往语音采集项目的参与者中招募合适的说话人来参加录音，这样自然产生时间间隔。LDC 选择的标准是该参与者在以往的项目中至少有过五次录音，同时这些录音的时间距 Greybeard 项目发起时至少有两年的间隔。录音模式类似 CSLU 与 MARP 的结合，说话人打电话到 LDC 的机器人操作终端，自动随机接通正在线上的另一说话人，同时操作终端给出一个建议的话题，两位说话人就该话题进行交谈，一般在 3 到 10 分钟之间。LDC 的机器人操作终端从 2008 年 10 月 7 日运行到 11 月 17 日，六周时间内 175 位说话人中至少有 100 位完成了 10 次电话语音采集，至少有 25 位完成了 20 次电话语音采集。这些说话人来自于 LDC 的过往电话语音采集项目，如：1990～1991 年录制的 Switchboard（Godfrey, 1992）、

1997 年录制的 Switchboard II、Switchboard Cellular（Miller *et al*., 2001）、2001～2003 年录制的 Mixer 1 & 2、2004 年录制的 Mixer 3（Cieri *et al*., 2006, 2007）。因此各说话人的时间间隔在 4～18 年不等。前后两批语音数据的时间间隔，每个说话人都各不相同，而且相对来说各批次内录音时间比较集中。

2.1.2.4　Trinity College Dublin Speaker Ageing (TCDSA) Database

这个数据库（Kelly and Harte, 2011; Kelly *et al*., 2012, 2013）是由爱尔兰都柏林三一学院（Trinity College Dublin）的电子与电气工程系和瑞士联邦理工学院（EPFL, Swiss Federal Institute of Technology Lausanne）从公开途径获取的语音资源中整理而来。数据库中共有 18 位说话人（性别平衡），每位说话人语音时间跨度为 30～60 年不等，其中各次语音间间隔 1～10 年不等。库中大多数语音资源来自于英国和爱尔兰的国家广播公司、英国广播公司 （BBC, British Broadcasting Corporation）以及爱尔兰国家电视和广播电台 （RTÉ, Raidió Teilifís Éireann）。此外还有部分公开语音样本来自于 YouTube 和弗吉尼亚大学 Miller 中心的总统演讲存档。

2.1.2.5　其他数据库

Markel 和 Davis（1979）曾经使用过一个时间跨度为三个多月的数据库，17 位说话人（11 位男性、6 位女性）。每人共有 10 次录制会话，间隔至少一周（一般为两到三周），通过录音机完成。每次会话为 15 分钟的一个访谈，话题不固定。

陆伟博士（2008）在研究中使用的是 YOHO Speaker Verification 英文数据库（Campbell and Higgins, 1994）（序列号为 LDC94S16），是录制于 1989 年的麦克风语音数据，共有 140 位说话人（32 位女性、108 位男性）。每位说话人在三个月的时间内在现实办公室环境下共录制 14 次，间隔一般为三天左右，每次录音会话由 24 个数字串构成（形式为 XX-XX-XX）。

日本电信电话株式会社 （NTT, Nippon Telegraph and Telephone）的说话人识别数据库 NTT-VR 也带有时变性质（Matsui and Furui, 1992），共有 35 位说话人（13 位女性、22 位男性）。每位说话人在 10 个月的时间内共录制 5 次会话，分别为 1990 年 8 月、9 月和 12 月以及 1991 年 3 月和 6 月。每次录制中说话人被要求以正常、快和慢语速分别说出日语句子，每句话大约 4 秒钟。

Beigi（2009, 2010）在研究中使用的数据库录制于 2007 年 8～12 月间，共有 22 位说话人。每人共有 3 次录制会话，间隔 1～2 个月左右。每次会话由多段 1 分钟长度的问题回答构成。

Lamel 和 Gauvin（2000）使用的是法国国家电信研究中心 （CNET, Centre national d'études des télécommunications，法语直译为 National Center for Telecommunication Studies）与机械与工程学科计算机实验室 （LIMSI, Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur，法语直译为 Computer Science Laboratory for Mechanics and Engineering Sciences）联合构思设计的数据库，语音通过法国电话网录制。数据库中共有 100 位说话人，每人在两年的时间内使用不同的电话设备在不同的地点总共产生 35 通电话语音。

此外，日本语音资源联盟 NII-SRC 中也有 AWA-LTR（AWA Long-Term Recording Speech Courpus）这样的时变数据库（NII, 2007a）。AWA-LTR 是由千叶大学（Chiba University）和大同大学（Daido University）组织录制的，6 位说话人周期性地录制 2～10 年，间隔一周左右，录音在隔音室完成，每次会话包括日语元音、词、四位数字串以及句子。但目前的发布版本中只有 1 位男性说话人 1 年的语音。

陈文翔等（2010）在研究中使用的数据采集于浙江大学 CCNT 实验室的声纹打卡系统（单振宇等, 2005）。他们选取了 2004 至 2009 年中长期使用打卡系统的 6 位实验室成员（2 位女性、4 位男性）的声纹打卡数据进行研究。这也是目前已知时变声纹资源中唯一基于中文语音的数据库。

2.1.3　构建合适的时变声纹库

从现有时变声纹资源的概述可以看出，大致存在以下几个问题：

CSLU、Greybeard 和 Lamel 等使用的数据库是通过远程电话语音方式采集的。这就不可避免地会受到其他环境因素的影响，诸如不可控的背景噪音程度、电话通信信道等。TCDSA 也存在类似的问题，其语音样本来源较复杂，包含了广播、电视采访以及演讲，样本间质量有很大差异。尽管作者称察看过语音样本的频谱内容，任何存在明显频率漂移（frequency artefacts）的样本都会被移除，但录音信道、设备及各种语音压缩格式之间的差异并非如此简单地体现在频率之上。

数据库时间跨度较短或者录制会话次数较少也是不利于时变研究的因素。如 Markel 和 Davis 使用的数据库、YOHO Speaker Verification 数据库、Beigi 使用的数据库等，时间跨度分别为 3 个月或 5 个月。NTT-VR 数据库时间跨度为 10 个月，但录制会话次数只有 5 次。

当然这与时变研究的特点有关，它需要反复组织相同的人员进行录音，这其中组织方面的困难可想而知。于是可以看到，一些时间跨度较长的数据库，其说话人数目就会急剧下降，如 AWA-LTR、TCDSA 等。对于说话人识别的研究来说，

几个人或者十几个人显然不是一个理想的规模。

关于说话方式方面，MARP、CSLU、Greybeard 和 TCDSA 等，都采用了自由交谈的对话方式（或访谈方式）。但在自由交谈对话方式下，说话人的情绪和参与度等都极容易受到搭档的影响，也许这些更易捕捉到的变化会掩盖掉部分时变因素。

另外一些数据库有时变性质，但并非专门为时变研究唯一目的而设计，例如 MARP 数据库前半部分的语音采集主要是为了不同语气或发音方式的研究。还有一些数据库，目前并没有发现其上的时变研究成果发表，如 CSLU。Greybeard 已经用到 NIST SRE 2010 评测之中，目前只对 NIST SRE 2010 的参与单位开放，但暂时还没有该数据库上结果的相关报告（NIST, 2010）。

正如绪论中所分析，一个合适的时变声纹数据库的缺失是制约此方面研究发展的关键。因此，作为 CCC 联盟的主要发起人，我们（清华大学语音和语言技术中心）试图录制一个合适的时变声纹数据库，录制完成后通过 CCC 平台进行发布，以供国内外有需要的研究机构、学术团体、大学以及其他科研单位研究使用。该时变声纹数据库命名为 Chronos（Chinese Reading lONgitudinal vOiceprint databaSe），下文将详细介绍 Chronos 的设计原则和具体实施方案。

2.2　Chronos设计原则

2.2.1　整体的设计原则

Chronos 的整体设计原则如图 2.1 所示。

研究说话人识别中的时变规律是本课题的唯一目标，因此，在 Chronos 的设计中，遵循的总原则就是——“尽最大可能”保证“时间”是整个语音采集过程中唯一变化的因素。

于是除了“时间”之外其他的因素，正如图 2.1 中所示，例如，语音采集所在环境和条件、语音采集所需要的软硬件设备等等，在整个时变语音的采集过程中应尽量保持不变。所以同一个录音室、同一套录音设备、同一套录音软件、维持在一个较低水平且大致相当差异不大的环境噪声和干扰，是时变声纹数据库建立的前提。

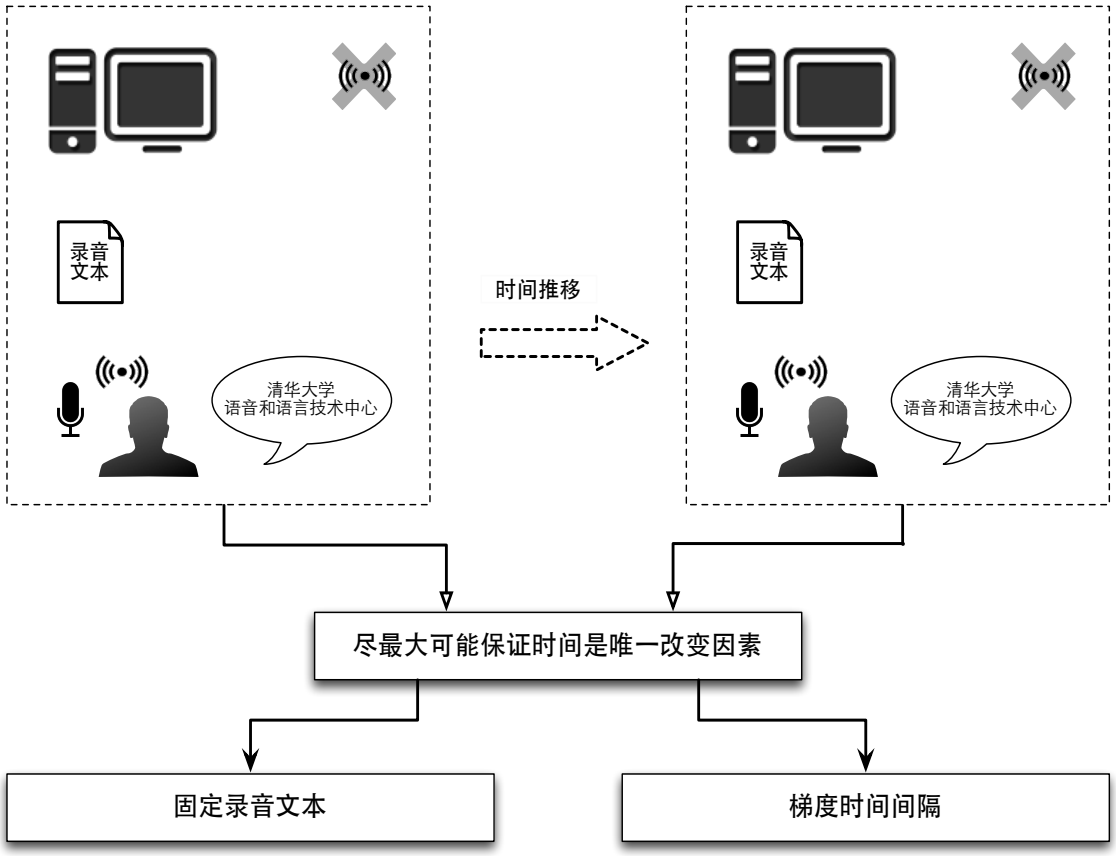

图2.1　Chronos设计原则

在这个设计总原则之下，有两个主要的因素需要考虑：录音提示文本的设计和语音采集时间间隔的设计（即各次录音会话的安排）。相关的设计原则详述如下。

2.2.2　固定的录音文本

前一小节里阐述了语音采集过程中可能的外部影响因素，以及对于这些因素的稳定性处理。而其实人本身又是一个更加复杂且随时变化的系统，其情绪和发音方式等特别容易受到外界的影响或干扰，这方面的稳定性是时变语音采集所面临的更大挑战。很多研究人员也探讨过不同情感类型下（黄挺, 2011; 单振宇, 2010），或者不同的语速（Nakagawa *et al*., 2004）、音量（Zhang and Hansen, 2007）等条件下，说话人识别系统性能的下降。显然对于时变课题而言，我们希望回避这些变化因素。这也是在之前的分析中，我们不主张使用自由交谈或者访谈方式进行语音采集的重要原因。

基于上述原因，我们在语音采集过程中采用了固定的录音提示文本，并要求

说话人以正常朗读的方式进行发音。当然固定的录音文本也进一步回避或者至少降低了语音的文本内容对于说话人识别性能的影响。跟其他语音类数据库一样，在录音文本的设计上，也需要充分考虑语音单元（如音节、声韵等）的覆盖率和均衡度。

Chronos 将录音文本组织成句子和孤立词的形式。这对应于说话人识别的文本无关和文本相关两种分类。文本无关的说话人识别，需要的有效语音较长，以句子的形式比较合适提取声纹；而文本相关的说话人识别，不需要很长的有效语音，一般的应用会选取两到五个字的孤立词以提取声纹。

2.2.3　梯度的时间间隔

从现有的时变声纹资源分析可以看出，关于录音会话间间隔的设置，各数据库的情况各不相同，大多只是规定在某个时间跨度内录制若干次会话，会话间间隔是否大致相等或者是否有其他考虑就不得而知了。另外，由于时变录音重复性的特点，为了得到一种可能的时变趋势而以一个长度固定的时间间隔进行十多次会话录制，一来人力物力耗费非常大，录音成本较昂贵，二来可能也没有必要。因此在 Chronos 数据库 3 年的录制时间跨度下，我们采用了梯度的时间间隔。

具体说来，最初几次录音会话间间隔大致为一周的时间，接下来几次录音会话间隔拉长至大约为一个月的时间，依次类推，随着 Chronos 录制的进行，各次录音会话间隔越来越长。梯度设计基于以下假设，说话人识别系统的性能在最初阶段下降比较剧烈，而随着时间的推移，性能下降并不会像初期时那么明显。当然，无论这个假设正确与否，我们都可以在后续的研究中很容易分析出不同时间间隔下说话人识别性能的变化规律。因此，最初的录音会话间我们选取了较短的时间间隔（一周或一个月），而之后的录音会话间逐步选取了更长的时间间隔（两个月、四个月甚至半年等）。

2.3　Chronos具体录制方案

以下分别从录音文本的设计、录音会话时间间隔的选取、说话人的选取和征募以及录音环境和软硬件设备的准备等方面来详细介绍 Chronos 的录制方案。

2.3.1　录音文本

录音文本由 100 个汉语句子和 10 个汉语孤立词组成。每位说话人在每次录音会话中都会朗读同样的文本。每个句子包含 8 到 30 个汉字不等，平均长度为 15；而每个词包含 2 到 5 个汉字，一次录音会话中每个词会被朗读 5 遍。其中 5 个词

与 100 个句子一样，在全部录音会话中保持不变；而另外 5 个词，每次录音会话都会替换为新的词语，为未来其他研究而保留。

汉语是一种由单音节构成的语言。每个音节由一个声母（Initial）和一个韵母（Final）构成。标准普通话中有 21 个声母、38 个韵母，而扩展的声韵母列表中还包含有 6 个零声母（Zero-Initial）（Lin and Wang, 1991; Li *et al*., 2001, 2004）。如表 2.1 所示。

表 2.1　扩展的普通话声韵母列表

类型	个数	列表
声母	21	b, p, m , f, d, t, n, l, g, k, h, j, q, x, z, c, s, zh, ch, sh, r.
韵母	38	a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn.
零声母	6	_a, _o, _e, _w, _y, _v.

语音中普通存在着连读的现象，声韵母的发音受上下文影响也较大，因此一般使用“di-IF”模型（Dobrisek *et al*., 1999）来考虑录音文本的覆盖率和均衡度，如表 2.2 所示。

表 2.2　普通话中的 di-IF 统计

类型	举例	个数
声母+韵母	q+ing	380
韵母+声母	ing+h	38*21 = 798
韵母+零声母	ua+_y	38*6 = 228
零声母+韵母	_w+uai	36
总计	-	1,442

候选语料是来自于人民日报的 6,000 个句子，利用鼓励低频单元 （ELF, Encouraging Low-Frequency units）算法（Li *et al*., 2003; Xiong *et al*., 2003）从中选取了 100 个句子。ELF 是一种替换式的语料选择算法，每次迭代时找出目标句子集合中对低频语音单元（这里使用的是 di-IF）贡献度最小的句子，然后使用当前候选句子集合中对目标集合中低频语音单元贡献度最大的句子来进行替换，迭代至不再有替换发生。这种替换方式可以有效地将候选语料中含有低频语音单元的句子挑选出来。最终的录音文本声学覆盖度如表 2.3 所示。

表 2.3　录音文本的声学覆盖度

类型	覆盖数目	总数	覆盖百分比（%）
声母	21	21	100
韵母	38	38	100
零声母	6	6	100
di-IF	1,183	1,422	82

2.3.2　时间间隔

Chronos 录制项目开始于 2010 年，预计在大约三年的时间内，每位说话人将完成 16 次录音会话。数据库采用了 5 种不同的时间间隔，分别是：一周、一个月、两个月、四个月和六个月（半年），如图 2.2 所示。

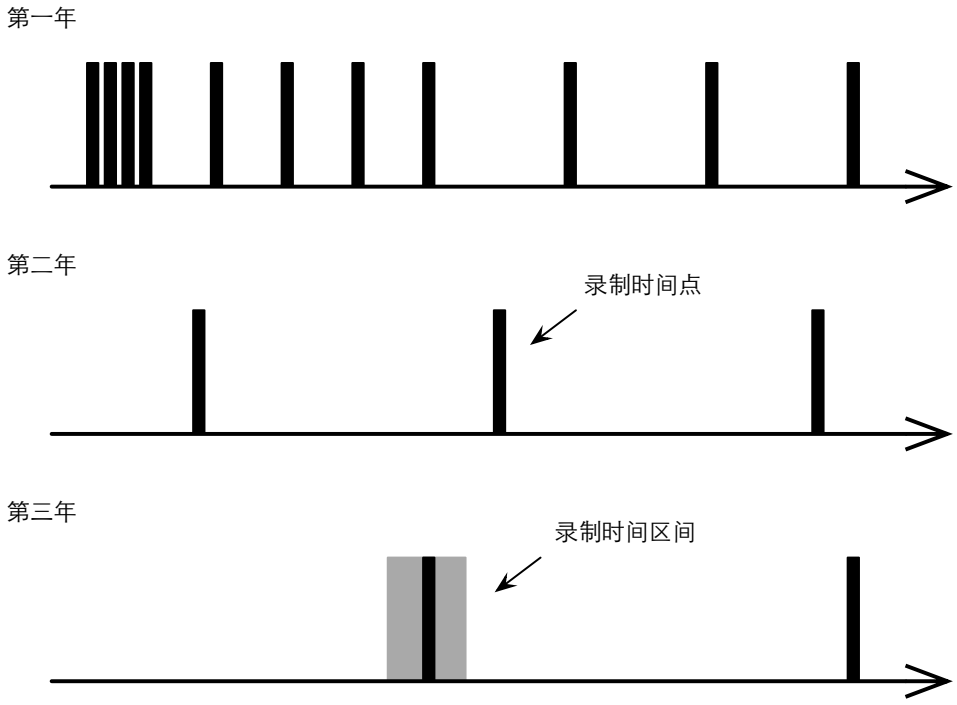

图2.2　Chronos录制时间间隔设计

图中展示了设计的完整 16 次录音会话在 3 年录音跨度中分布的具体时间点（图中黑色代表录制时间点），但考虑到在实际录制过程中很难要求所有说话人在指定的某一天全部完成会话录制，因此录制时间点就灵活地变为了区间形式（图中灰色代表录制时间区间），一般前后相差一周左右。

2.3.3　说话人

绪论中 1.4.1.2 节曾经介绍了生理解剖学上关于不同年龄阶段的发声系统变化情况的研究。可以看出，儿童期、青春期以及老年期都是生理器官变化比较剧烈的阶段，而变声期过后至中青年时期虽然基音频率依然随时间变化呈现出一个缓慢下降的趋势，但相对而言仍是一个比较稳定的阶段。这个年龄段的人群也是现实中说话人识别应用主要面向的群体，研究其声纹的时变规律更有现实意义。因此，我们将 Chronos 的目标说话人首先定位于这个年龄段。

考虑到时变声纹录制的长期反复性，要求参与项目的说话人能在 3 年的时间跨度内保持基本固定的居住地，因此我们从清华大学校内征募了 60 位大学生作为 Chronos 的说话人（性别平衡）。说话人的院系和出生年份分布如图 2.3 所示。

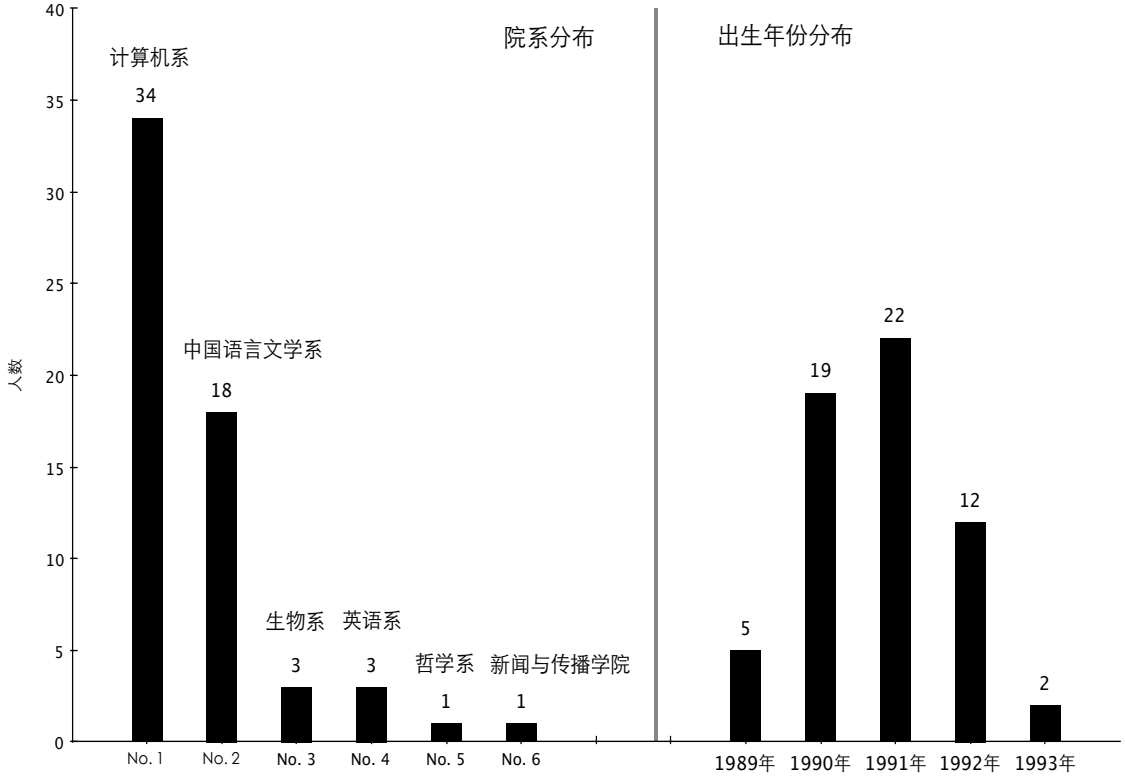

图2.3　Chronos中说话人的院系和出生年份分布

尽管他们来自于国内不同的省份和地区，但都可以讲比较流利的普通话。而录音时间间隔的设计也特意避开了暑假和寒假等大多数人可能不会留在校园里的情形。

2.3.4　录音环境等

实验室的一个隔间专门用作录音室。实验室环境保证了没有突发噪声，背景噪声也维持在一个较低的水平。在第一次录音会话时，工作人员会告诉参与录制的说话人如何操作录音软件及其他注意事项。说话人被要求以正常的语速朗读出录音软件界面上依次出现的每个句子和词语；此外，录音软件有音量控制模块，当说话人的音量超出预设的接受范围时，软件会提示说话人音量过高或过低，并要求说话人重新录制当前内容。大多数说话人可以在 25 分钟内顺利完成一次录音会话。

语音信号将同时以 8kHz 和 16kHz 两种采样率、16 比特精度进行录制。

由于第一次录音会话主要以掌握录音软件使用和熟悉录音流程为目的，因此之后的研究和实验中并没有使用第一次录制的语音数据，而是以第二次及之后录制的语音数据为主。

2.4　Chronos上的时变表现

本节从语音的频谱特征、声纹特征以及说话人识别系统性能等方面分别探讨数据库上的时变表现。

2.4.1　频谱特征

利用 PRAAT 工具（Boersma, 2002）对频谱进行了分析，包括基音频率和共振峰等的变化。图 2.4 展示了编号为“007”的说话人（女性）跨度半年的 6 次“法（fa）”（三声）的发音频谱对比。此图为 PRAAT 工具截图，其中上面一行为 6 次发音的波形图，中间和下面两行黑白部分为频谱图，中间一行的蓝色间断线为各次发音的基音频率，而下面一行的红色间断线为各次发音的共振峰情况。从频谱图中可以看出，基音频率和共振峰随时间变化并没有非常显著的差异，尽管第 6 次与第 1 次相比，基音频率和第四共振峰有下降趋势，这与生理解剖学上的结论相似。二者的变化更接近于一定范围或者区间内的随机抖动。

2.4.2　声纹特征

2.4.2.1　声纹特征的选取

这里的声纹特征指的是说话人识别任务中通常会用到的声学特征。尽管基音频率和共振峰反映了人本身的生理结构特点，携带了一定的说话人个性信息，但在实际的说话人识别研究中最常用的是倒谱系数特征。

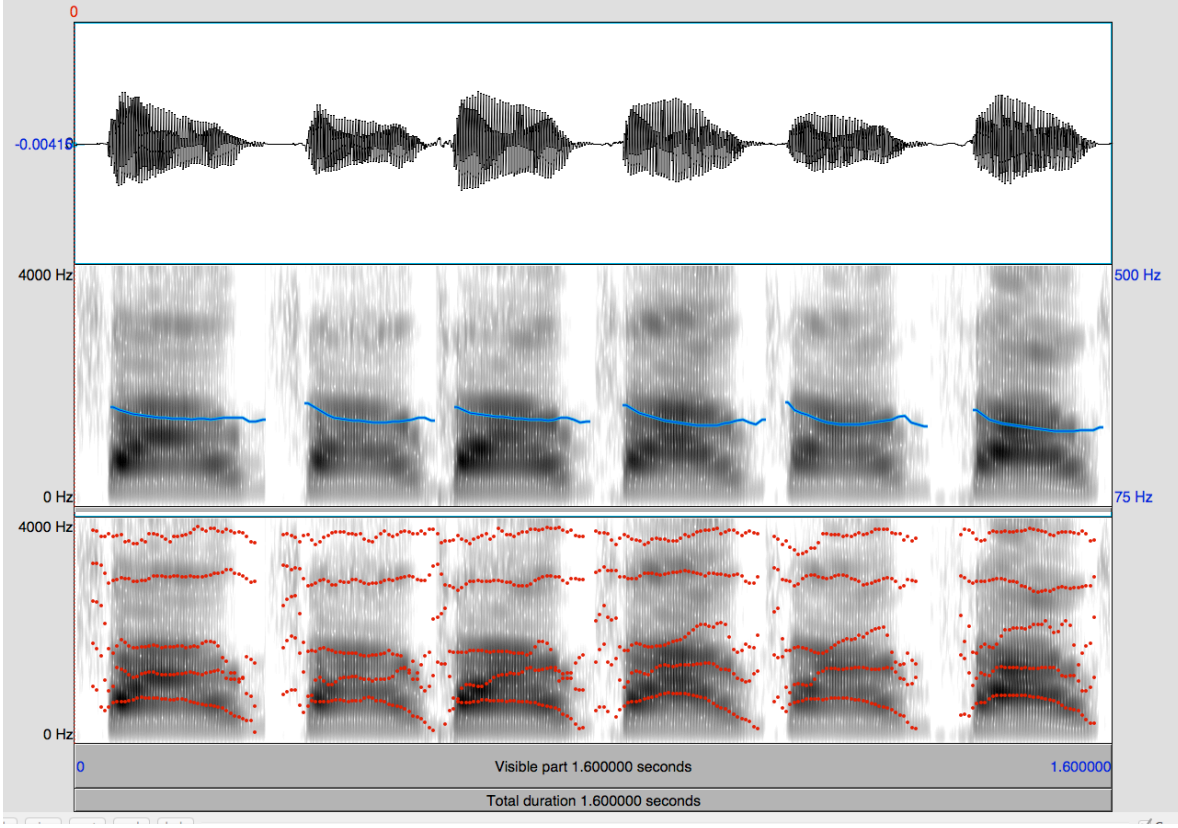

图2.4　编号为“007”的说话人跨度半年的6次“法（fa）”的发音频谱图对比

倒谱系数特征各个维度间相互独立，降低了统计建模处理的难度，而且特征中还包含了全部频带的信息，使其应用更加广泛和灵活。而倒谱系数特征中，又以 MFCC 为最常用。

Mel 频率是基于人耳的听觉特性而提出的，与赫兹 Hz 代表着客观音高不同，它是主观音高的单位。从图 1.4 可以看出二者的非线性对应关系，MFCC 特征实际上是加强了低频。而低频部分通常被认为与语音内容更相关，因此 MFCC 一直是语音识别中的经典特征。但对于说话人识别，几十年的研究中 MFCC 也一直是主流特征。不少研究者提出，理论上讲应该对语音内容相关的部分进行抑制的说话人识别任务，仍旧使用 MFCC 特征是否合适（Lu and Dang, 2007, 2008）。当然不可否认的是，MFCC 特征大多数情况下在识别系统中性能非常稳定，这也是它被广泛采用的一个重要原因。

因此，这里依然选取 MFCC 特征来进行时变分析，目标是考察相隔一段时间的两段语音，它们的声纹特征参数，是否会随着时间的推移而逐渐拉大“距离”。

2.4.2.2　相关度的计算

选用的是 Pearson 提出的相关系数（Correlation Coefficient）（Rodgers and

Nicewander, 1988; Stigler, 1989; Goh *et al*., 2007）来度量两段语音的特征参数之间的“距离”。对于两个随机变量 X 、Y 来说，其相关系数 ρ 计算公式如下：

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y}. \qquad (2\text{-}1)$$

其中，$\operatorname{cov}(X,Y)$ 为两个随机变量间的协方差，而 σ 、E 、μ 分别为随机变量的标准差、期望以及均值。

倒谱系数特征的提取是以帧为单位的，若计算相关系数时也以帧作为单位，那么计算随机变量 X 、Y 的联合分布就会遇到困难。因为即使同一位说话人发同样的音，每次发音的帧数也不可能完全相同。而考虑到 Chronos 的录音文本是固定的，即所有说话人所有录音会话中，其语音内容都是完全相同的（即音素的种类、数目和出现顺序都是固定的）。因此我们选取了音素作为计算相关系数的单位。

计算两次录音会话间的相关系数时采取了如下步骤：

（1）音素时间对齐。根据对应的录音文本中的音素信息，利用自动语音识别中的强制对准（Forced Alignment）技术来获得所有语音数据中各个音素的起止时间点；具体地说，依赖于 HTK（Young *et al*., 2002）提供的 HVite 工具（-a 参数），利用正确文本（-I 选项给定标准输入）自动建立起没有歧义的搜索网络，其中搜索得分最高的时间点序列就是强制对准的结果；

（2）以帧为单位提取语音的 MFCC 特征；

（3）根据步骤（1）中得到的各元素起止时间点信息，得到以音素为单位的特征参数。通常某一音素对应着 N 帧数据，则将相应的 N 帧倒谱系数特征进行算术平均，作为该音素的特征参数；

（4）得到公式中所要求的各个统计量的分布（期望、方差等），计算出相关系数。由于倒谱系数一般为多维，假设各维之间相互独立，得到各维的相关系数，最后求其算术平均，作为两次录音会话间的最终相关系数。

在 Chronos 上，我们以第二次录音会话为基准，分别计算了接下来的 8 次录音会话与第二次之间的相关系数。MFCC 采用了 16 维系数。相关系数变化趋势如图 2.5 所示。

从图 2.5 可以看出，随着时间推移，MFCC 特征参数的相关系数整体而言有下降趋势。可见，单纯从声纹特征的角度来讲，时变现象依然是客观存在的。

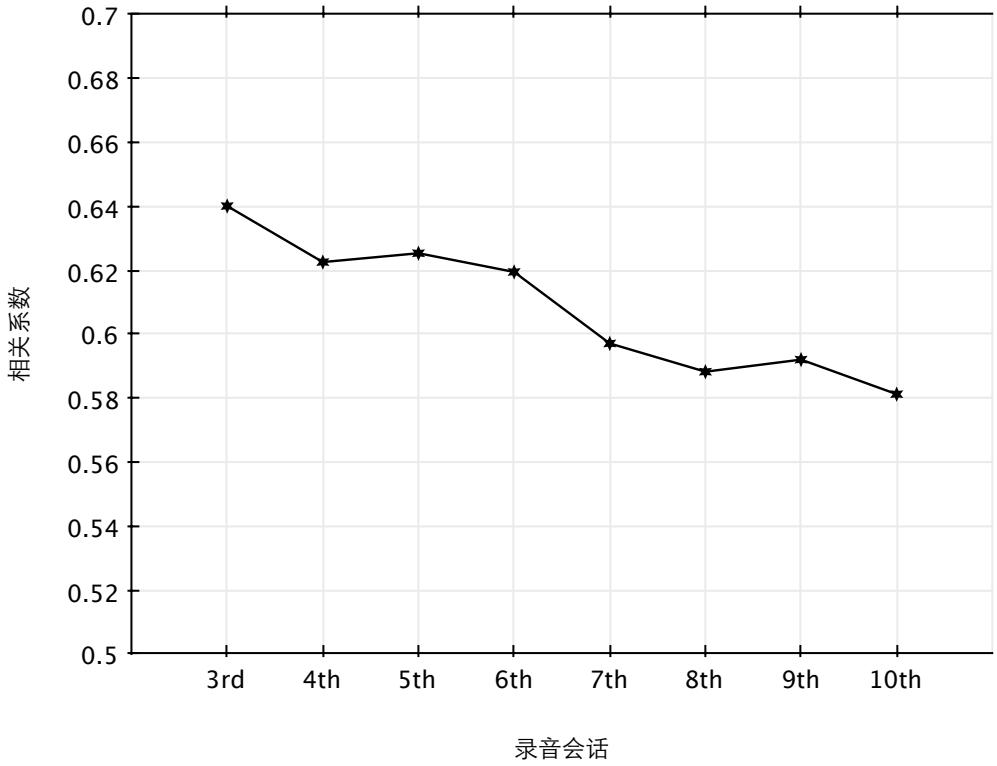

图2.5　后续录音会话分别与第二次之间的相关系数变化趋势

2.4.3　系统性能

采用经典的 MFCC 特征和 GMM-UBM 模型来构建说话人识别系统，以考察系统的时变表现。MFCC 选取了 16 维系数加上 16 维一阶差分，GMM-UBM 模型选用了 1024 混合。UBM 模型利用实验室现有的麦克风语音训练而成。训练说话人模型的语音来自于第二次录音会话，从每人的 100 个句子中随机选出 3 个句子，拼成长约 10 秒钟的一段语音。除了训练语音之外，每次录音会话的所有句子都用于测试（第二次至第十四次），平均每句时长约为 2 到 5 秒钟。不同录音会话的等错误率 EER 变化曲线如图 2.6 所示。

可见随着训练数据与测试数据之间时间间隔越来越长，说话人识别系统的性能也逐渐恶化。尽管后一次并不是绝对会比前一次 EER 高（比如第九次的 EER 要略高于第十次），但整体的趋势依然十分明显。

图 2.6 以录音会话编号作为横坐标，但由于 Chronos 中采用的是梯度的时间间隔，横坐标与时间并非线性对应关系。为了更方便看到随着天数递增，系统性能的变化趋势，在图 2.7 中采用了距第二次录音会话的天数作为横坐标。

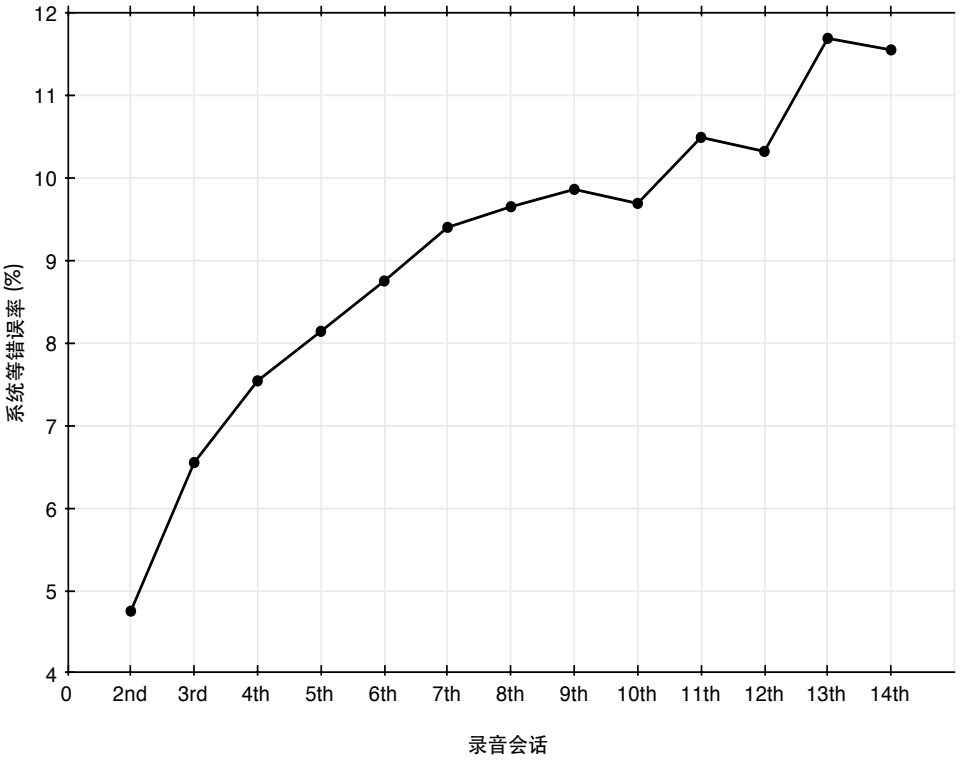

图2.6 不同录音会话的等错误率变化曲线（训练数据来自第二次）

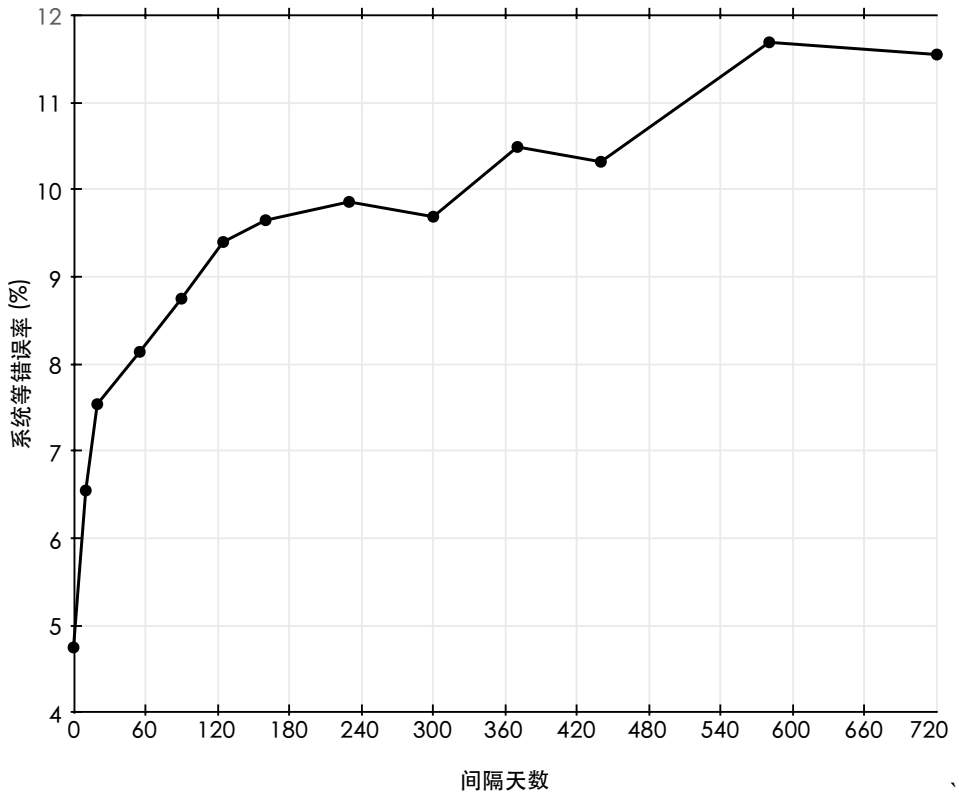

图 2.7 不同录音会话的等错误率变化曲线（以间隔天数作为横轴）

图 2.7 有力地支持了 Chronos 在设计时间间隔时的假设：说话人识别系统的性能在最初阶段下降比较剧烈，而随着时间的推移，性能下降并不会像初期时那么明显。

以上是训练数据来自第二次录音会话的情况，当训练数据分别来自第三次和第七次录音会话时，系统性能变化曲线如图 2.8 所示。两条整体趋势类似“V”字型的曲线再次体现出了时间变化对说话人识别系统性能的影响。

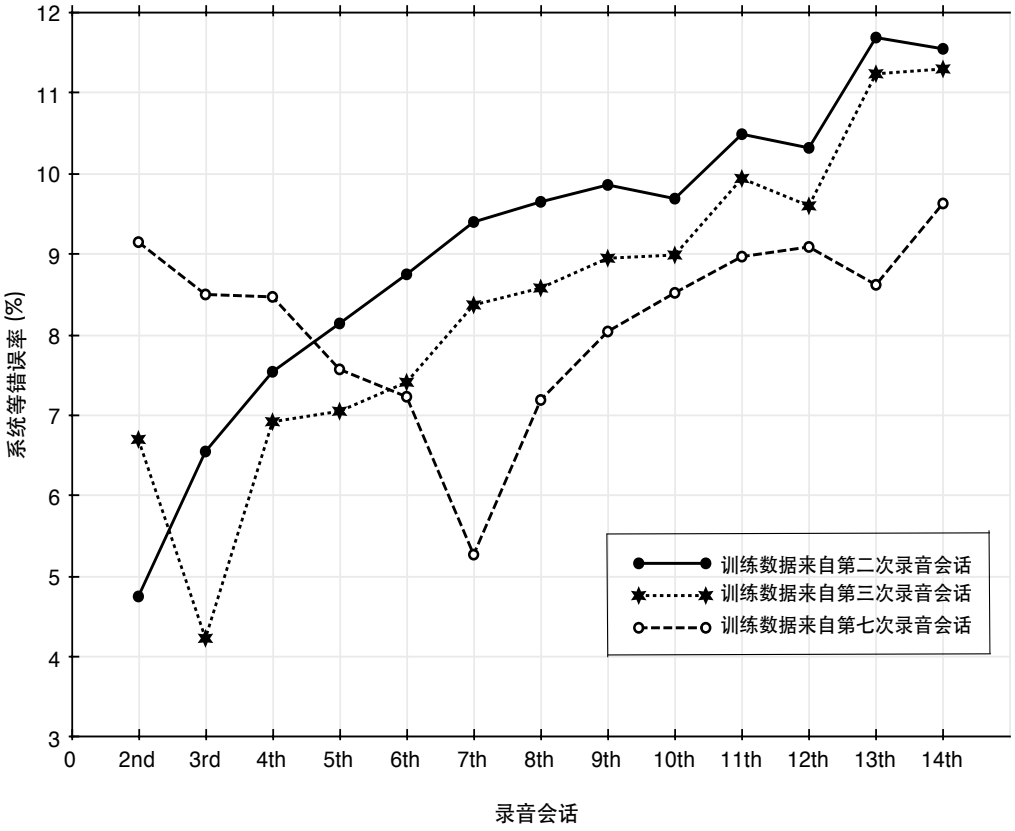

图 2.8 不同录音会话的等错误率变化曲线（训练数据分别来自第三、四次录音会话）

2.5 小结

本章从语音领域的资源联盟建设和时变声纹数据资源现状入手，提出录制一个合适的时变声纹数据库的目标，希望通过 CCC 平台发布，以供国内外有需要的研究机构、学术团体、大学以及其他科研单位研究使用。接着阐述了构建 Chronos 时变声纹数据库的原则——除了时间这个变化因素之外尽最大可能保持其他因素不变，以及两个具体的原则——固定的录音文本和梯度的时间间隔。然后详细说明了 Chronos 的各项具体录制方案，包括录音文本的组织、时间间隔的设计、说话

人的征募以及录音环境和条件等。最后从具象的频谱特征、声纹特征相关度以及系统识别率的变化等方面展现出了 Chronos 上的时变现象。

第3章　基于以F-ratio为中间准则计算频带区分度的时变鲁棒特征提取算法

3.1 频带区分度

3.1.1　频带区分度的概念

作为人类最自然和有效的交流和沟通方式，语音中传递着多种类型的信息，例如，与语音内容相关的语义信息、与说话人身份相关的个性信息、语速音量信息、说话人情感信息以及说话场景背景信息等。简言之，一段语音传达了“什么”人在“什么”环境下以“什么”方式、“什么”状态表达了“什么”内容。表达这些“什么”信息的各类信号以一种复杂的方式互相影响且交织叠加在一起，就构成了语音信号。因此，语音研究中的各个应用，如说话人识别、语音识别、情感识别、场景识别等，本质上就是从语音信号这个复杂的整体中剥离出相应类型的信息的过程。

然而各类信息在语音信号频带间的分布被普遍认为是不均匀的（Zhou *et al*., 2011）。一般认为与语义内容相关的信息（即音素区分度较高的信息），如前三个共振峰，存在于 200Hz 到 3,000Hz 这一中低频区域内，因而这一区域对于语音识别来说更为重要（Rabiner and Juang, 1993; Stevens, 1998）。而声门和梨状涡等被认为与说话人个性信息紧密相关的信息则分别主要反映于 100Hz 到 400Hz 的低频区域以及 4,000Hz 到 5,000Hz 的高频区域内（Lu and Dang, 2007, 2008）。从这些实验语音学的研究可以看出，不同的频带对于不同的应用任务有着不同的贡献，于是不同的应用任务对于不同频带的关注度也不尽相同。

本文中将不同频带对于不同应用任务的贡献程度，称之为该频带对于特定任务的区分度。

3.1.2　时变说话人识别中的频带区分度

很多研究人员确认了说话人个性信息在不同频带间的分布也是不均匀的（Lu and Dang, 2007, 2008）。于是对于时变的说话人识别这一具体任务而言，一种可能的假设是这种与时间相关的说话人个性信息在语音中不同频带间的分布也是不均匀的。而与时间相关的个性信息分布较多的频带，对时变的说话人识别而言自然是不利的。它更应该关注的频带是那些分布有较多稳定的说话人个性信息的频带，而对于分布有较多时间相关信息或者较少说话人个性信息的频带都应给予较少的

关注度。

时变说话人识别以寻找时变表现稳定的声纹特征为目标，该任务包含两个层次：捕捉到说话人个性信息且同时忽略掉与时间有关的信息。 因此既考虑频带对于说话人个性信息的区分度，同时也考虑其对于时间相关信息的区分度，综合二者即可以得到时变的说话人识别任务的频带区分度。该区分度与前者（说话人个性信息）正相关，而与后者（时间相关信息）负相关。

明确了各频带对于时变说话人识别的整体区分度，就可以依据此区分度对各频带进行相应的强调和弱化——在特征提取的过程中，强调那些对说话人个性信息表现出较高区分度、同时对时间相关信息表现出较低区分度的频带，而弱化表现反之的频带，从而提高说话人识别系统的时变鲁棒性。

3.1.2　频带区分度的确定准则

如何确定各个频带在时变说话人识别任务上的区分度，是本文要解决的一个核心问题。经典的确定方法大概可以归为以下几类：（1）利用由先验知识等方式获得的参数，如 Mel 频率弯折方式即是利用生理学实验确定的等响度曲线；（2）利用预先设计的特征选取准则直接对特征相关参数进行选取；（3）直接依据识别系统的性能来调整特征参数等。通过生理解剖学的先验知识获取的频率参数一般具有较好的推广性，但在某些问题上往往缺少可用的先验知识，时变说话人识别即是如此。 因此本文详细探讨了两种确定准则，一是基于频带能量和 F-ratio 的准则，一是性能驱动准则，分别对应于后两类确定方法。

本章将主要探讨基于频带能量和 F-ratio 的确定方法，而第 4 章和第 5 章则主要探讨两种不同的性能驱动准则。

3.2　基于频带能量和F-ratio的准则

3.2.1　以F-ratio为频带区分度的中间准则

F-ratio，一般用以判断哪种统计模型更适用于当前的数据集合（这里指的是广义的统计模型）。对于模式识别问题来说，这一准则经常被用来判断哪种建模方式更为有效，或者哪种特征提取方式更为合适。F-ratio 定义如下：

$$F\_ratio = \frac{\text{between-group variability}}{\text{within-group variability}}. \qquad (3\text{-}1)$$

费舍尔（Ronald A. Fisher）于 1920 年首次使用了方差相比的方式作为统计量，

F-ratio 也得名于此（Lomax and Hahs-Vaughn, 2007）。从定义可以看出，如果某一种统计模型（如特征）在特定数据集合的某一种划分（分类，即 grouping）方式上拥有较高的 *F-ratio* 值，则说明该模型对该数据划分方式具有较大的类间距离，或者较小的类内距离。即通过这一特征提取方式，可以使得该数据划分方式拥有较大的类间离散度和较小的类内离散度。这就意味着对于这一特定的数据划分方式而言，这种特征提取方式能较好的捕捉到对分类更有价值的信息，即前文所提到的更高的区分度。在语音领域中，很多研究人员使用 F-ratio 来挑选合适特征（Wolf, 1972），Lu 和 Dang（2007, 2008）也曾利用 F-ratio 准则来判断不同频带对于一般说话人识别任务的区分度。

对于一般的说话人识别任务，显然语音数据按不同说话人进行划分即可，而在时变说话人识别任务中，情况更为复杂，存在两种不同的数据划分方式，如图 3.1 所示。

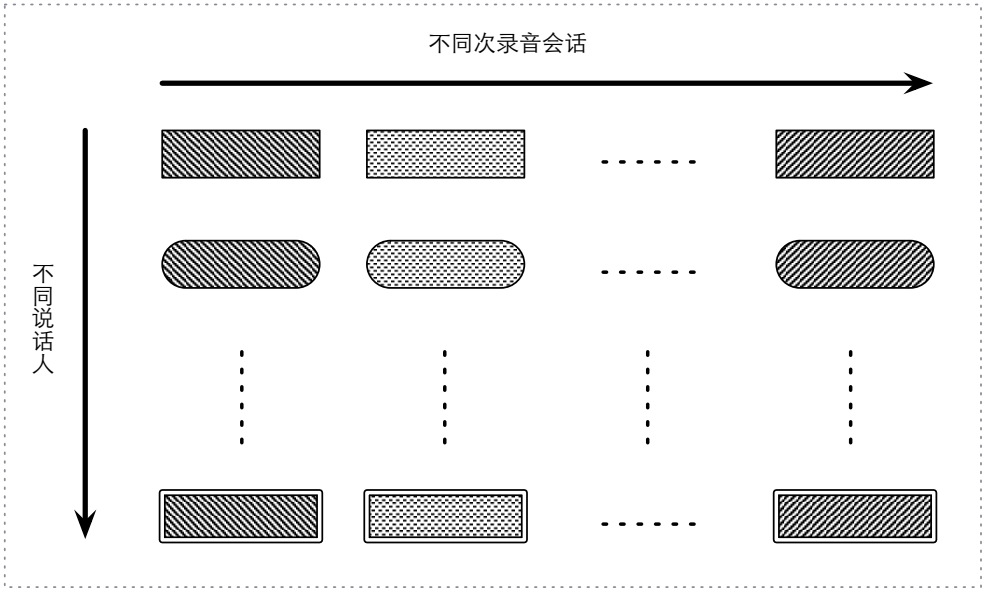

图3.1　两种不同数据划分（说话人和录音会话）示意图

上图中不同的形状表示不同的说话人，不同的填充纹理则表示不同次的录音会话（本文称之为 session）。所有时变语音数据可组织成如图所示的二维形式。于是，对应着两种数据划分方式，时变说话人识别中就存在着两种 *F-ratio*。将每个说话人的语音数据作为一类，可以得到与说话人个性信息的区分度相关的 *F-ratio*，称之为 *F_ratio_spk*，这也是一般的说话人识别任务中的 *F-ratio*；而将每一次录音会话的语音数据作为一类（即按 session 分类），可以得到与时间信息的区分度相关的 *F-ratio*，称之为 *F_ratio_ssn*，这是特定的时变说话人识别任务中特

有的 *F-ratio*。

3.2.2 频带能量作为参数

由绪论中图 1.3 可以看出，在倒谱系数类特征的提取过程中，经过 DFT 变换后的频带能量谱扮演着很重要的角色。于是本文选用频带能量来作为确定频带区分度的一种参数。具体做法如图 3.2 所示。

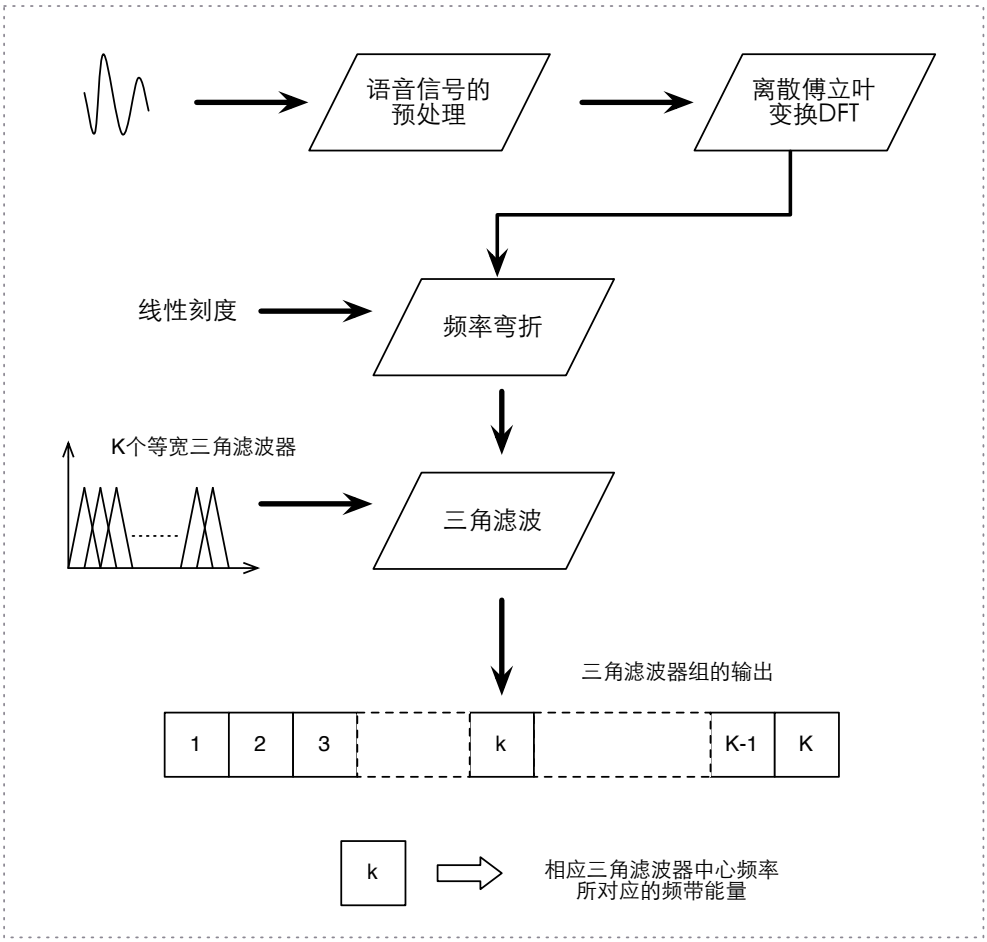

图3.2　频带能量参数计算过程示意图

图 3.2 中，线性刻度保证了后续操作依然在赫兹域完成，赫兹域 K 个等宽度的三角滤波器以其各自中心频率作为中心将整个频域范围划分成了 K 个频带。而三角滤波器的输出即为对应的频带能量参数。在之后的 *F-ratio* 的计算中，使用的即是对数能量。

3.2.3 两种F-ratio的计算

假设整个频域范围被分成 K 个频带。从 Chronos 中选取 M 个说话人、S 次录

音会话的语音数据来计算两种 *F-ratio*。那么，对于某一频带 *k*，在计算 *F_ratio_spk* 时，首先将语音数据按录音会话分成 *S* 批次，各批次内以说话人分类，计算出一个 *F-ratio* 值，然后对这 *S* 个 *F-ratio* 作几何平均即得 *F_ratio_spk*；同样的道理，将语音数据按说话人分成 *M* 组，各组内以录音会话分类，计算出一个 *F-ratio* 值，然后对这 *M* 个 *F-ratio* 作几何平均即得 *F_ratio_ssn*。

与说话人个性信息的区分度相关的 *F_ratio_spk* 计算如下：

$$F\_ratio\_spk_{s,k}=\frac{\sum_{i=1}^{M}\left(\mu_{i,s,k}-\mu_{s,k}\right)^2}{\sum_{i=1}^{M}\frac{1}{N_{i,s}}\sum_{j=1}^{N_{i,s}}\left(x_{i,s,j,k}-\mu_{i,s,k}\right)^2}. \tag{3-2}$$

其中，$F\_ratio\_spk_{s,k}$ 表示频带 *k* 在第 *s* 次录音会话的 *F-ratio* 值，$x_{i,s,j,k}$ 是说话人 *i* 在第 *s* 次录音会话语音数据第 *j* 帧的频带 *k* 对应的对数能量，$N_{i,s}$ 是说话人 *i* 在第 *s* 次录音会话语音数据的总帧数，而 $\mu_{i,s,k}$ 和 $\mu_{s,k}$ 是相应的均值（期望），计算如下：

$$\mu_{i,s,k}=\frac{1}{N_{i,s}}\sum_{j=1}^{N_{i,s}}x_{i,s,j,k}. \tag{3-3}$$

$$\mu_{s,k}=\frac{1}{M}\sum_{i=1}^{M}\mu_{i,s,k}. \tag{3-4}$$

于是对于每一个频带 *k*，与说话人个性信息的区分度相关的 *F_ratio_spk* 由公式（3-5）得到。

$$F\_ratio\_spk_{k}=\left(\prod_{s=1}^{S}F\_ratio\_spk_{s,k}\right)^{\frac{1}{S}}. \tag{3-5}$$

类似地，与时间信息的区分度相关的 *F_ratio_ssn* 计算如下：

$$F\_ratio\_ssn_{i,k}=\frac{\sum_{s=1}^{S}\left(\mu_{i,s,k}-\mu_{i,k}\right)^2}{\sum_{s=1}^{S}\frac{1}{N_{i,s}}\sum_{j=1}^{N_{i,s}}\left(x_{i,s,j,k}-\mu_{i,s,k}\right)^2}. \tag{3-6}$$

其中，$F\_ratio\_ssn_{i,k}$ 表示频带 *k* 在说话人 *i* 的 *F-ratio* 值，$x_{i,s,j,k}$ 和 $N_{i,s}$ 定义同前，而 $\mu_{i,k}$ 的计算如下：

$$\mu_{i,k} = \frac{1}{S}\sum_{s=1}^{S}\mu_{i,s,k}. \tag{3-7}$$

于是对于每一个频带 k，与时间信息的区分度相关的 F_ratio_ssn 由公式（3-8）得到。

$$F\_ratio\_ssn_k = \left(\prod_{i=1}^{M} F\_ratio\_ssn_{i,k}\right)^{\frac{1}{M}}. \tag{3-8}$$

3.2.4 整体区分度的定义

这样经过 3.2.3 节的一系列计算，每一频带 k 都对应着一对 *F-ratio* 值：$F\_ratio\_spk_k$ 和 $F\_ratio\_ssn_k$，分别对应着说话人个性信息和时间相关信息。对于时变说话人识别而言，频带的整体区分度应该与 $F\_ratio\_spk_k$ 正相关（强调说话人个性信息），而与 $F\_ratio\_ssn_k$ 负相关（弱化时间相关信息）。因此我们定义频带 k 的整体区分度 $Discrim\_F\_ratio$ 如下：

$$Discrim\_F\_ratio_k = \log\left(\frac{F\_ratio\_spk_k}{F\_ratio\_ssn_k}\right). \tag{3-9}$$

这样针对时变说话人识别任务的特点，就得出了以频带能量为参数同时利用 F-ratio 作为中间准则的频带整体区分度曲线。

3.3 时变鲁棒性算法

如前所述，明确了各频带对于时变说话人识别这个特定任务的整体区分度，就可以依据此区分度对各频带进行相应的强调和弱化，从而提高说话人识别系统的时变鲁棒性。

频带的强调和弱化，其实是通过改变来自不同频带的信息在所提取的特征中所占权重来对特征进行的调整。常见的调整方法包括：（1）对离散傅立叶变换 DFT 后得到的所有样本点，即所有离散频率分别加权（Miyajima *et al*., 2001）；（2）改变某些特定频带附近输出滤波器的密度，即频率弯折的方法；（3）改变各个滤波器输出的强度，即滤波器输出加权的方法（Miyajima *et al*., 2001; Biem *et al*., 2001）。所有离散频率分别加权的方法对区分度的刻画要求更加精细，参数亦非常庞杂，因此本文对于时变鲁棒性算法的研究，主要关注于三角滤波器组一前一后的两处设置——滤波前的频率弯折和滤波后的输出加权，即后两种调整方法。

3.3.1　频率弯折

3.3.1.1　总体思想

所谓频率弯折，实质是通过调整各频带弯折后的分辨率来达到强调或者弱化不同频带的效果。频率弯折之后在变换域的等宽三角滤波器组设置相当于频率域上三角滤波器的不等宽非均匀分布。梅尔 Mel 刻度就是一种典型的频率弯折方式。取 MFCC 特征时 Mel 域和其等价的赫兹域的三角滤波器组设置对比如图 3.3 所示。

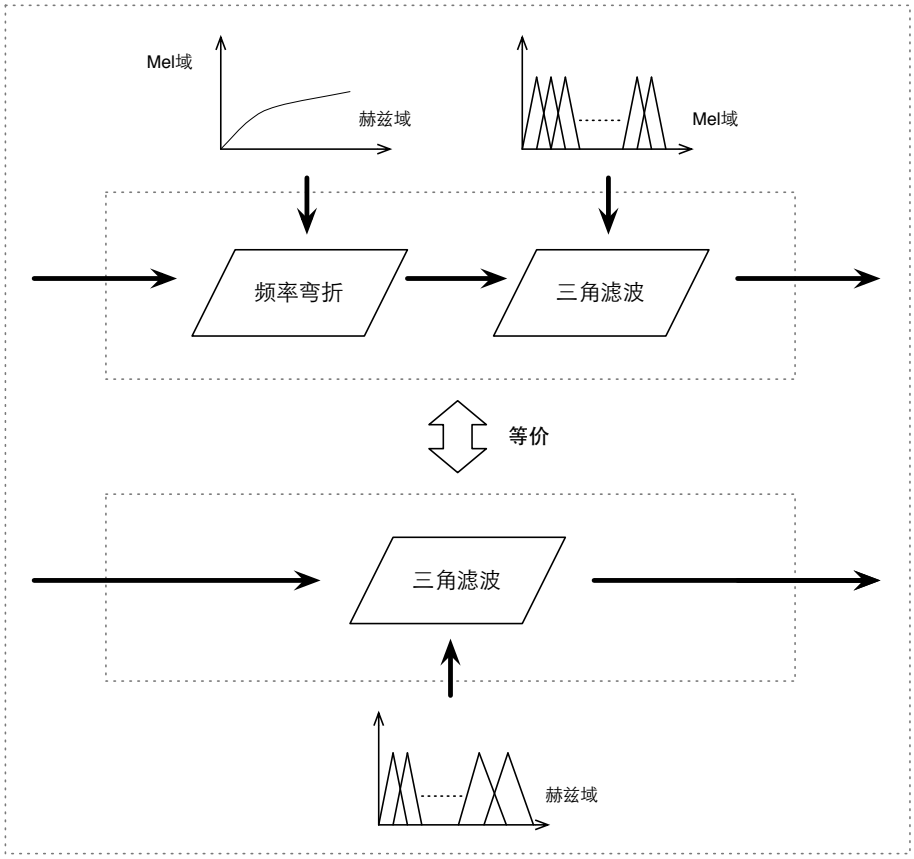

图3.3　提取MFCC特征时Mel域和赫兹域的三角滤波器组设置示意图

由图可见，Mel 刻度的频率弯折相当于改变了赫兹域三角滤波器组设置的疏密分布。低频区域的三角滤波器设置更密集，宽度更窄，通常称之为分辨率越高，特征提取时可从该区域提取更多的信息；而高频区域的设置更稀疏，宽度更宽，通常称之为分辨度越低。

更具体地，以 8kHz 采样的语音信号提取 MFCC（Mel 刻度）特征为例，假设三角滤波器个数为 30，这里以三角滤波器的频率带宽（Hz）来代表其所对应的中心频率的分辨率，那么 MFCC 在不同频带的分辨率如图 3.4 所示。

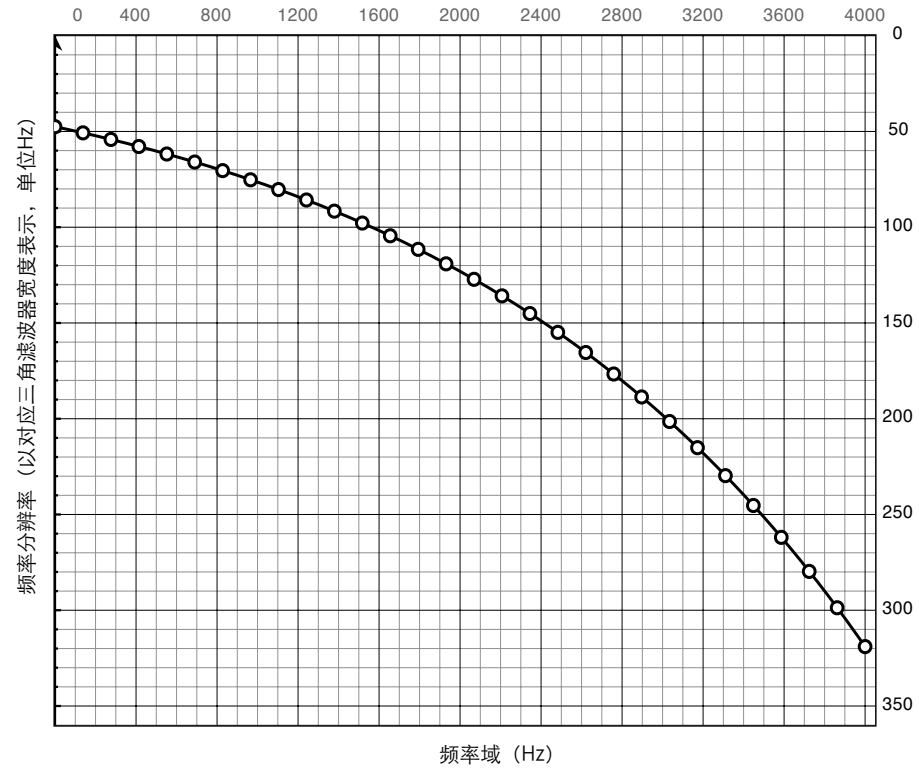

图3.4　MFCC特征提取时不同频带的分辨率

说话人识别领域不少研究人员将寻找一个最优的频率弯折曲线作为研究目标（Zhou *et al*., 2011）。本章也是利用类似的思想，通过确定各频带对于时变说话人识别这一特定任务的区分度，来试图找到对这一任务更有效的频率弯折方式，从而由语音信号中提取出更为稳定的说话人个性信息，以提高说话人识别系统的时变鲁棒性。

3.3.1.2　具体算法

于是，按照之前计算得出频带的整体区分度设计频率弯折曲线。曲线设计的准则详述如下。

以一对相邻的频带 *k* 和 *k-1* 为例。我们假设频带的整体区分度 $Discrim\_F\_ratio_k$ 是 $Discrim\_F\_ratio_{k-1}$ 的 2 倍，则两个频带在频率域和变换域的相对关系如图 3.5 所示。

而所有频带基于整体区分度的频率域和变换域的相对关系，相当于对图 3.5 作一全频域范围的扩展。于是整体的频率弯折曲线反应在坐标平面上就是一条由 *K* 段斜率不等的直线连结而成的折线段。

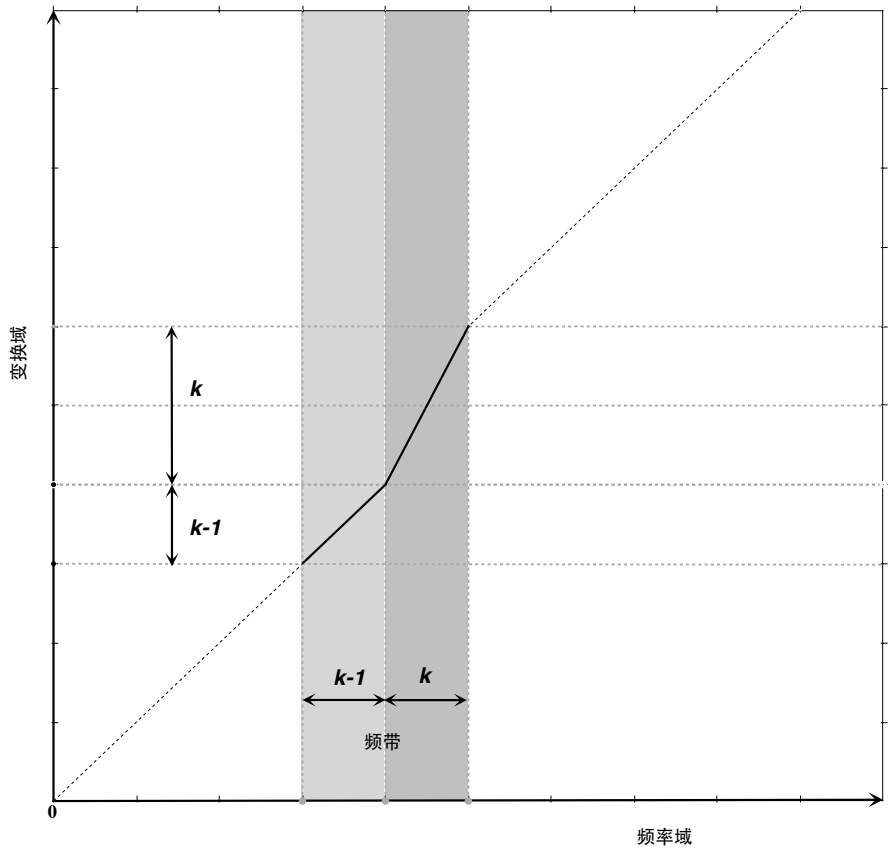

图3.5 两频带基于整体区分度的频率域和变换域相对关系示意图

根据弯折曲线可以确定频率域（赫兹域）与变换域的关系，将频率范围映射到变换域内，然后通过在变换域内设置等宽的三角滤波器组来实现对所有频带的加强或弱化处理。当然，如图 3.3 所示，这样的操作其实等同于在频率域直接设置不等宽的三角滤波器组来实现。三角滤波器间的相对宽度（分辨率）同样可以通过区分度来计算。这里不作赘述。

3.3.2　滤波器输出加权

与频率弯折相比，滤波器输出加权是更加直接的一种调整方式。在各滤波器输出生成倒谱系数之时，依据整体区分度进行直接的加权，从而调整相应频带对于倒谱系数的贡献度。

按照图 3.2 中所示步骤，得到对数能量谱，记为$S(k)$，而加权后的对数能量谱记为$Weighted_S(k)$，计算如下：

$$Weighted\_S(k) = Discrim\_F\_ratio_{k+1} \cdot S(k). \qquad (3\text{-}10)$$

于是，根据倒谱系数$Cepstrum(n)$的计算过程（见绪论中图 1.3），其后的 DCT

变换公式如（3-11）所示。

$$\begin{aligned} Cepstrum(n) &= \sum_{k=0}^{K-1} Weighted\_S(k)\cdot\cos\left(\frac{\pi n(k+0.5)}{K}\right) \\ &= \sum_{k=0}^{K-1} Discrim\_F\_ratio_{k+1}\cdot S(k)\cdot\cos\left(\frac{\pi n(k+0.5)}{K}\right). \end{aligned} \tag{3-11}$$

3.4　实验

3.4.1　实验设置

从 Chronos 中选取了第一次录音会话之后的十三次会话数据（第二次至第十四次）。每一次录音会话的语音均被等分为两部分，一部分作为开发集，用来计算频带的整体区分度，另一部分用做训练和测试。说话人模型的训练语音来自于第二次录音会话，由随机选取的 3 个句子组成，长度大约为 10 秒；测试时使用全部十三次录音会话的语音数据，每个句子为一条测试语音，长度大约为 2 到 5 秒。

实验中使用 8,000 Hz 采样语音，全频域范围（0 至 4,000 Hz）被分成为 30 个频带，即系统中使用的三角滤波器个数为 30。

基线系统选用了说话人识别研究中经典的 MFCC 特征加 GMM-UBM 模型的设置。其中 MFCC 特征使用了 16 维 MFCC 参数再加上 16 维一阶差分；GMM-UBM 建模使用了 1024 混合的高斯模型。本文之后用到的所有倒谱系数特征均采用同一配置——16 维倒谱系数加 16 维一阶差分。其中 UBM 模型利用实验室原有的麦克风语音训练而成。

3.4.2　整体区分度

在进行整体区分度的计算时，将全频域范围（0 至 4,000 Hz）均等分为 30 个频带。

根据公式（3-2）计算每次录音会话中各个频带的 $F\_ratio\_spk_{s,k}$，得到各次会话的说话人个性信息区分度曲线。图 3.6 展示了相隔大约半年左右的五次录音会话（分别为第二次、第八次、第十一次、第十三次和第十四次，即 s 分别为 2、8、11、13 和 14）的该区分度曲线。

从图 3.6 可以看出，尽管具体数值上或多或少有所差异，但这五次录音会话的说话人个性信息区分度曲线的趋势基本一致。

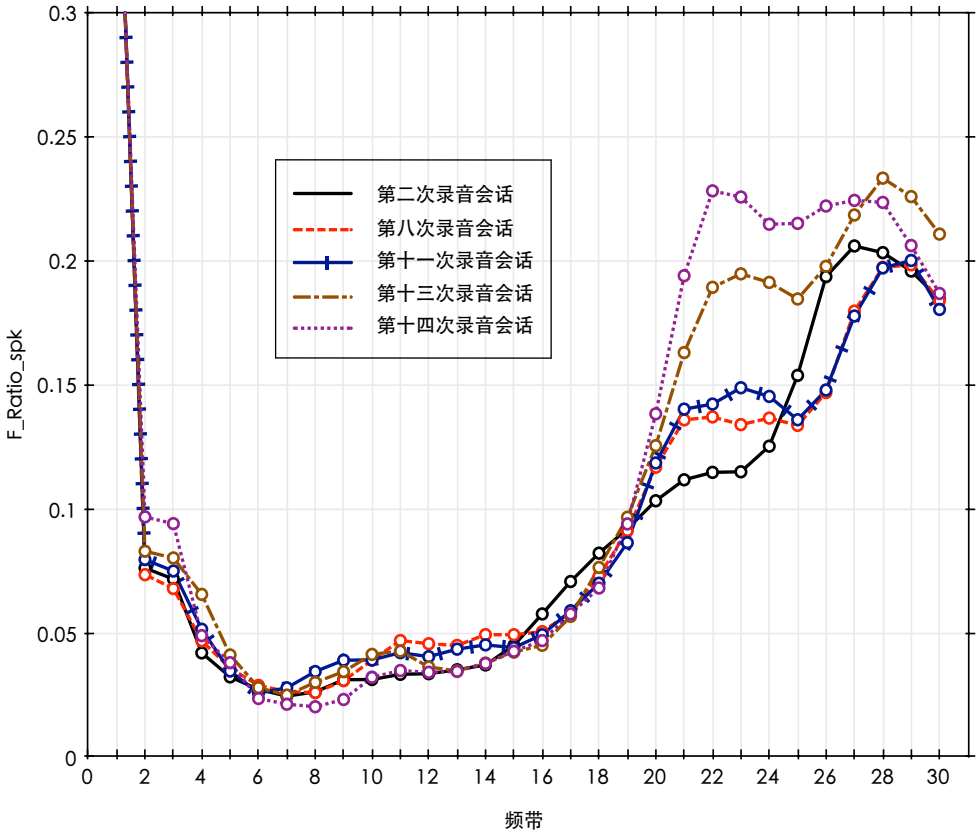

图3.6 间隔大约半年的五次录音会话的对应说话人个性信息区分度曲线示意图

在低频部分，五次录音会话的区分度之间相差不大，然而到了高频部分，尤其是在 2,500~3,500 Hz 的范围内，各次的区分度之间相差较大，并且大致呈现了随着时间推移，说话人个性信息的区分度增大的趋势。这些变化在某种程度上印证了说话人个性信息，即通常所谓声纹，存在着时变的部分，毕竟在数据库录制时已尽量保持各次录音会话间其他因素的稳定。

根据公式（3-5）和（3-8）计算得到的 $F\_ratio\_spk_k$ 曲线和 $F\_ratio\_ssn_k$ 曲线如图 3.7 所示。

由图 3.7 可见，两条曲线表现各异。其中，$F\_ratio\_spk_k$ 的曲线在频带 5 和频带 16 之间（约 600~2,000 Hz 的范围内）基本维持在一个相对稳定的数值，没有明显的变化；但那之后，曲线迅速攀升，并在频带 21 和频带 28 达到两个局部峰值。而 $F\_ratio\_ssn_k$ 的曲线在频带 6 之后以一个几乎稳定的速度上升，攀升过程中同样有两个局部峰值，但分别出现在频带 11 和频带 26。

最终，根据公式（3-9）计算得到的整体区分度曲线如图 3.8 所示。

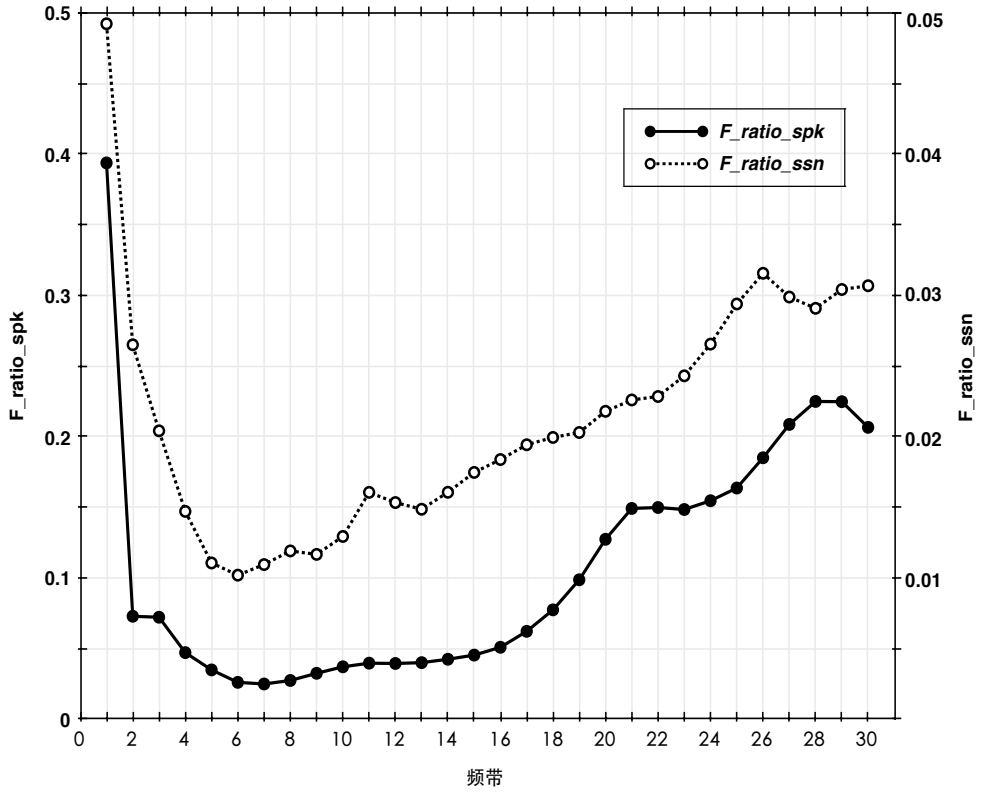

图3.7 两种区分度（*F_ratio_spk*和*F_ratio_ssn*）曲线示意图

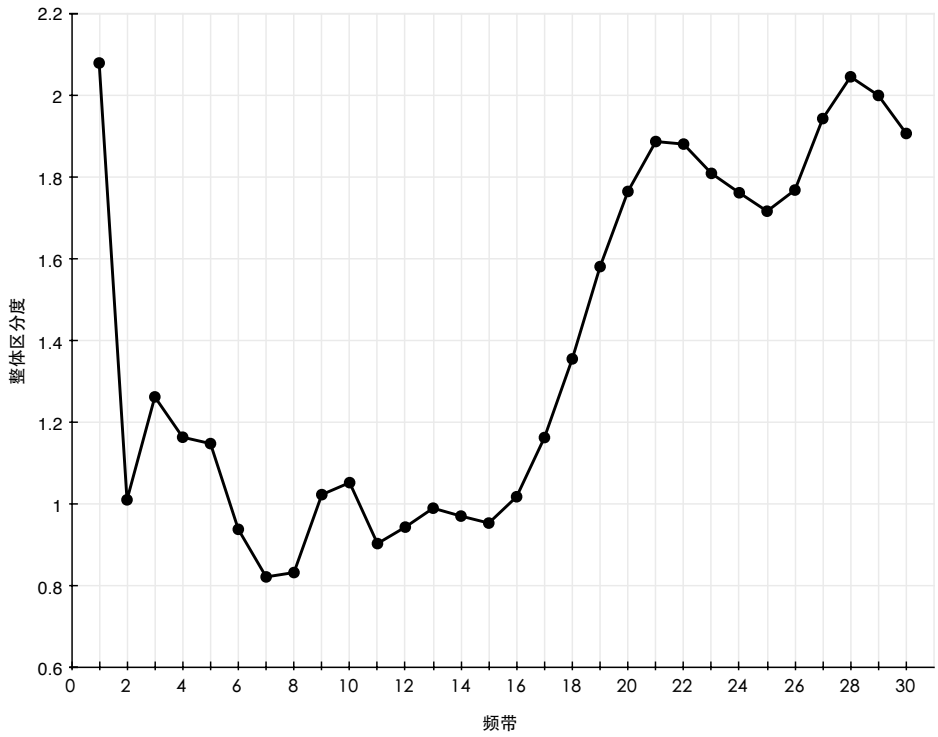

图3.8 以基于频带能量的F-ratio为中间准则计算得到的频带整体区分度曲线

图 3.8 中的频带整体区分度曲线是图 3.7 中两条曲线的一个折中：高频部分应当被强调，但不应加强至 $F\_ratio\_spk_k$ 曲线的那种程度，因为高频部分同样对应着较高的 $F\_ratio\_ssn_k$；低频部分亦有类似的表现。

3.4.3　实验结果

基于上述频带整体区分度曲线，对之前提到的两种鲁棒性算法：频率弯折和滤波器输出加权，分别提取倒谱特征（16 维倒谱加一阶差分）进行了说话人确认的实验，分别记为 Warping_F_ratio 和 Weighting_F_ratio，并与基线系统的 MFCC 特征进行了比较。各次录音会话的等错误率表现详见表 3.1。

表 3.1　三种倒谱特征方法的比较（等错误率 EER，%）

录音会话	MFCC	基于频率区分度的时变鲁棒性算法	
		Warping_F_ratio	Weighting_F_ratio
2nd	4.5	4.0	4.1
3rd	6.4	6.1	6.5
4th	7.4	6.6	7.2
5th	8.1	7.5	7.8
6th	8.7	7.2	7.8
8th	9.3	8.5	9.1
9th	9.9	7.9	8.7
10th	9.6	8.1	8.5
11th	10.0	8.9	10.0
12th	9.7	8.8	9.7
13th	11.1	9.7	10.1
14th	11.0	9.4	9.7

从表 3.1 中可以看出，频率弯折的方法在每一次录音会话数据中的表现都要优于基线的 MFCC，而滤波器输出加权的方法在大部分录音会话数据中的表现优于 MFCC，只有第三次录音会话是个例外，MFCC 表现稍微好一些。

表 3.2 从这十三次录音会话的等错误率的均值和标准差这两个统计量的角度进一步比较了三种特征方法。

表 3.2　三种倒谱特征方法的比较（等错误率的均值和标准差，%）

倒谱特征方法	等错误率		下降率	
	均值	标准差	均值	标准差
MFCC	8.80	1.86	--	--
Warping_F_ratio	7.77	1.54	11.70	17.20
Weighting_F_ratio	8.32	1.67	5.45	10.22

从均值和标准差两个统计量的表现可以看出，基于频带整体区分度的两种时变鲁棒性算法整体来说在各录音会话数据上的说话人识别等错误率表现均优于MFCC；尤其是标准差这个统计量，它一定程度上反映了等错误率的时变特性，而两种算法相对于基线系统均取得了超过 10%的下降，在一定程度上缓解了说话人确认系统性能的时变恶化程度。关于时变说话人识别的性能综合评价指标将在第 4 章中详细阐述。

而两种鲁棒性算法对比来看，在以基于频带能量的 F-ratio 作为中间准则来计算频带整体区分度的前提下，频率弯折的方法在所有录音会话数据中的表现都要强于滤波器输出加权的方法，这说明后者可能需要一种更为复杂、精细的程序（或步骤）来进行强调或弱化。更进一步的讨论将在第 5 章中详细阐述。

3.5　小结

本章从频带区分度谈起，提出了在时变说话人识别的任务中频带整体区分度的概念，并简要阐述了区分度的确定准则。而本章着重从基于频带能量和 F-ratio 的准则入手，详细介绍了频带整体区分度的计算。确定了频带的整体区分度之后，就要在特征提取时强调那些对说话人个性信息区分度高且对时间相关信息区分度低的频带，而弱化表现反之的频带。本章探讨了与滤波器组设置相关的两种强调和弱化的鲁棒性算法：频率弯折和滤波器输出加权。最后通过实验中反映的现象分析了区分度曲线，并对基线系统和本章提出的两种鲁棒性算法从等错误率及其统计量的角度进行了性能比较。

第4章 基于性能驱动的频率弯折方法的特征提取算法

4.1 性能驱动准则

4.1.1 基于频带能量和F-ratio准则的局限

第 3 章中探讨了利用 F-ratio 为中间准则并以频带能量为参数来进行整体区分度的计算，但在实际系统中频带能量毕竟只是特征提取过程中的一个中间变量，它的区分度还会受到模型设计、参数优化等因素的影响，因而基于频带能量和 F-ratio 准则的区分度高与系统实际性能好是否总是一致，这也是一个问题。

Lei 和 Lopez 曾经重复过 Lu 等人计算说话人个性信息相关 F-ratio 的实验（Lei and Lopez, 2009），得出了相似的 *F-ratio* 变化趋势，他们与本文第 3 章中的具体计算方法略微有所不同，但得到的 *F_ratio_spk* 曲线的走向趋势一致，高频区域的 *F-ratio* 值远高于低频区域的 *F-ratio* 值。这也意味着与 Mel 刻度不同，说话人识别更应加强高频区域，这也与普遍认为的高频区域含有更多说话人个性信息的说法一致。但在随后的 MFCC、LFCC（线性频率倒谱系数）以及 a-MFCC（即 anti-MFCC，频率弯折方式刚好与 Mel 刻度方式相反）对比实验中，MFCC 和 LFCC 性能相差无几，但是 a-MFCC 并没有取得预期中的性能优势。

此外，从 F-ratio 准则的相关计算公式可以看出，*F-ratio* 的高低从某种意义上确实说明了两个频带之间的强弱对比关系，但是这种强调或者弱化的倍数关系并不一定可以直接用 *F-ratio* 的值来刻画。

因此，一种更为直接的办法是利用实际得到的说话人识别系统的性能来评估各频带的区分度，即性能驱动的原则。本章和第 5 章均是在这一原则之下，分别从频率弯折和滤波器输出加权两个方面分别来探索针对时变说话人识别的性能驱动的具体准则。

4.1.2 针对频率弯折的性能驱动准则

本章的核心是针对频率弯折方式的性能驱动准则，很自然的，频带的整体区分度是由单独对该频带作频率弯折而得到的系统性能来决定的。具体作法是，对于某一指定的频带，保持其他所有频带的分辨率不变（即维持线性刻度），而唯独加强该频带（加密三角滤波器的设置），利用这样的频率弯折方式提取倒谱特征，并设计出一个说话人识别系统，而该系统在时变数据上的性能就作为该指定频带的整体区分度指标。在这个过程中，针对频率弯折的方式，就完成了一次性

能反馈。

于是，该性能驱动准则（一次反馈）的总体框架如图 4.1 所示。

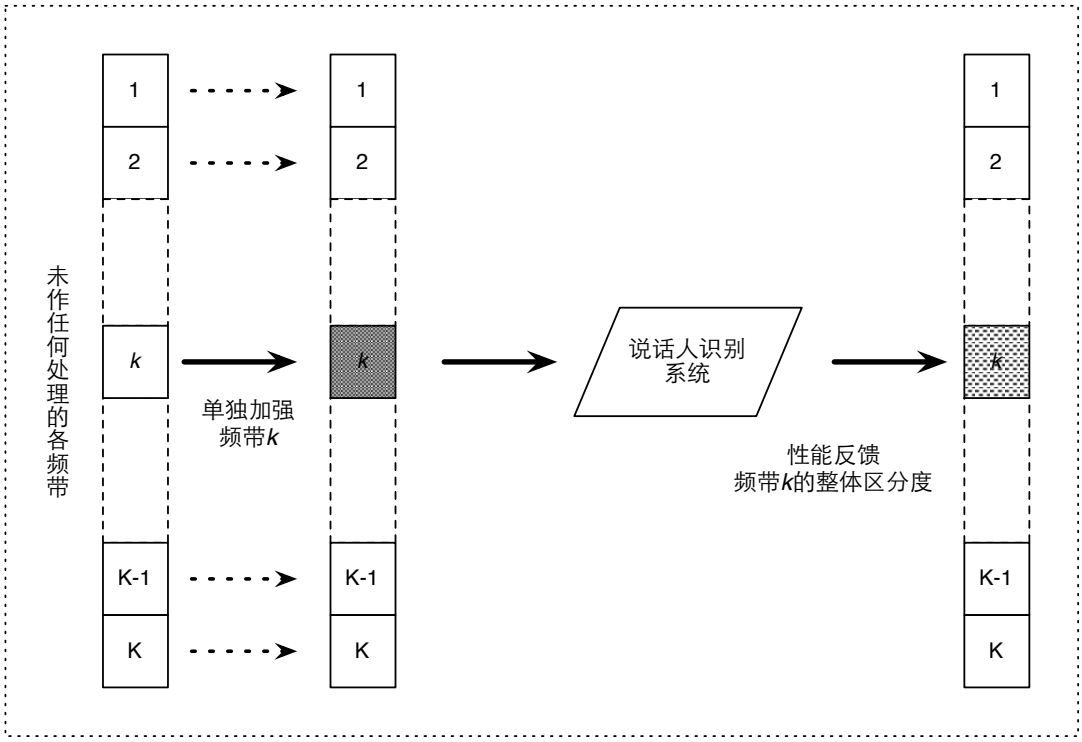

图4.1 性能驱动准则的总体框架示意图

性能驱动准则中有两个关键问题：怎样单独加强某一频带和选取怎样的系统性能指标，之后两节将有详细阐述。

4.2 频带的单独加强

在线性刻度的基础上单独加强某一频带，实质是单独提高该频带的分辨率，使得该频率域三角滤波器设置更密，较之第 3 章中所提全频域的强调和弱化更为简单。

以频带 k 加强两倍为例，经过频率弯折，频率域和变换域的关系如图 4.2 所示。可见频带 k 区域内频率弯折曲线的斜率由线性刻度的 1 变为了 2，这也意味着该区域内三角滤波器宽度是其他区域的一半。

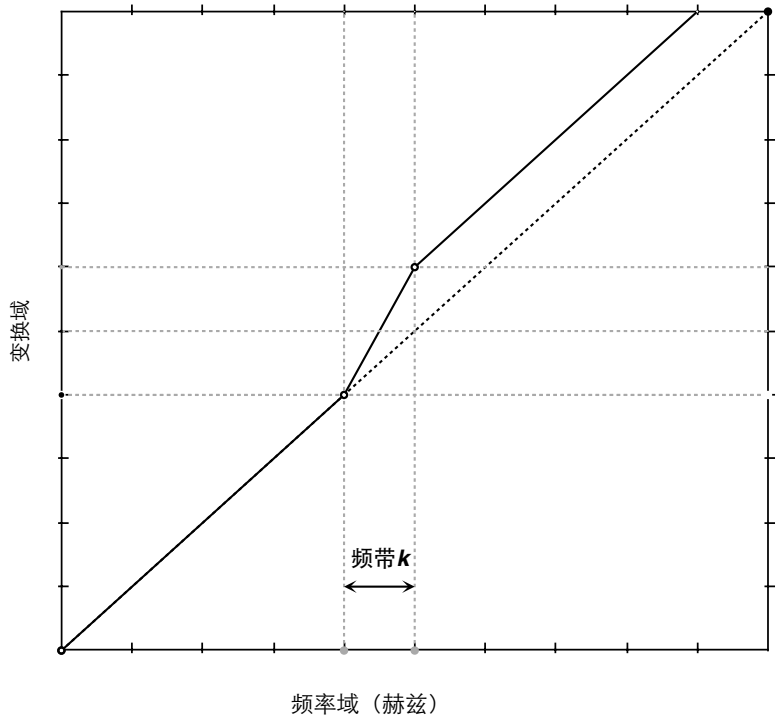

图4.2 单独加强频带k（两倍时）的频率弯折示意图

4.3 系统的性能指标

4.3.1 性能评价指标

4.3.1.1 说话人确认任务

以一个实际的说话人确认系统为例，错误接受率 FAR 和错误拒绝率 FRR 是一对重要的性能指标参数。它们跟说话人确认系统的阈值选取紧密相关。当阈值逐渐调高时，系统将真实说话人判定为假冒者的概率会增大，即错误拒绝率会升高，另一方面，系统将假冒说话人判定为真实说话人的概率相应减小，即错误接受率会降低，系统的安全性较好；相反，当阈值逐渐调低时，错误接受率会升高，而错误拒绝率会降低，系统面临的风险较高。因此二者是相互矛盾的关系。综合考虑这两个指标参数，人们习惯上用二者数值相等时的错误率作为衡量系统整体性能的指标，称之为等错误率 EER。

而在研究时变问题时，由于存在多次录音会话，因此随着时间的推移，系统将得到一系列 EER。这些 EER 的均值反映了系统在说话人确认任务上的平均性能，均值越小表示平均性能越好；而这些 EER 的标准差则反映了系统的时变鲁棒性，标准差越小表示系统性能随时间的变化越小，鲁棒性越好。二者综合反映了一个

时变说话人确认系统的性能。

因此，很自然地，对于一个说话人确认系统的整体时变鲁棒性能而言，EER 的均值减小，整体性能会有所改善，同样地，EER 的标准差减小，整体性能亦有改善。综合考虑这两方面的因素，本文在研究时变说话人确认系统整体性能时，选取了 EER 的均值和标准差的乘积来作为整体时变鲁棒性的一个综合评价指标，其数值越小代表该说话人确认系统的整体时变鲁棒性越好。

4.3.1.2 说话人辨认任务

对于说话人辨认任务，一般选用前 N 选的正确率（Top-N Accuracy）作为辨认系统性能的评价指标，而这其中又以首选正确率（即 N 为 1 的情况）在实际应用中最有价值。

与确认任务类似地，对于时变问题而言，由于存在多次录音会话，因此随着时间的推移，说话人辨认系统将得到一系列首选正确率，也即得到一系统首选错误率（100% -首选正确率）。这些首选错误率的均值反映了系统在说话人辨认任务上的平均性能，均值越小表示平均性能越好；而这些首选错误率的标准差则反映了系统的时变鲁棒性，标准差越小表示系统性能随时间的变化越小，鲁棒性越好。二者综合反映了一个时变说话人辨认系统的性能。因此，同样道理，对于一个时变说话人辨认系统，综合考虑两方面的因素，可以选取首选错误率的均值和标准差的乘积来作为整体时变鲁棒性的一个评价指标，其数值越小代表该说话人辨认系统的整体时变鲁棒性越好。

4.3.2 频带整体区分度

依然以说话人确认任务为例，根据系统的整体时变鲁棒性指标来确定频带的整体区分度，单独加强后进行识别操作得到的 EER 均值和标准差的乘积较小的频带，我们认为具有更高的整体区分度。

假设整个频域范围被分成 K 个频带，如图 4.1 所示。从 Chronos 时变声纹数据库中选取 M 个说话人、S 次录音会话的语音数据来计算区分度。那么，对于某一频带 k，单独加强后可得到一组 EER 值。定义性能驱动准则下频带的整体区分度 *Discrim_Pfm_Drvn* 如下：

$$Discrim\_Pfm\_Drvn_k = \log\frac{1}{\mu_k\sigma_k}. \qquad (4\text{-}1)$$

其中 μ_k 和 σ_k 为对应的 EER 均值和标准差，计算如下：

$$\mu_k = \frac{1}{S}\sum_{s=1}^{S}\xi_{s,k}. \tag{4-2}$$

$$\sigma_k = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\xi_{s,k}-\mu_k\right)^2}. \tag{4-3}$$

其中$\xi_{s,k}$表示单独加强频带 k 时录音会话 s 所对应的等错误率。

这样针对时变说话人识别任务的特点，就得出了利用性能驱动准则的频带整体区分度曲线。

4.4 实验

4.4.1 实验设置

实验数据的选取和划分、三角滤波器的个数、基线系统以及倒谱系数的配置与第 3 章的实验设置完全一致，详见 3.4.1 节，这里不再赘述。不同之处在于，3.4.1 节中选取的开发集数据用以训练 F-ratio 准则相关的参数进而得到频带的整体区分度；而本节选取的开发集数据进行一系列说话人识别实验从而根据实际性能得到频带的整体区分度。

4.4.2 整体区分度

4.4.2.1 加强倍数的确定

性能驱动准则的做法是在线性刻度的基础之上单独加强某一频带来考察其反馈区分度，那么要加强多少倍比较合适，这是首要解决的一个问题。线性刻度可以看作是单倍加强，在开发集上测试了双倍加强和三倍加强的效果。每一加强倍数对于每一频带都会有一组对应的等错误率的均值和标准差，因此在比较不同加强倍数的性能时，采用了对所有频带各组等错误率的均值和标准差再取平均（%）的评价指标，分别记为$AVG(\mu_k)$和$AVG(\sigma_k)$。性能如下表所示：

表4.1 加强倍数的性能对比

倍数	$AVG(\mu_k)$	$AVG(\sigma_k)$
单倍	8.02	1.60
双倍	7.61	1.49

三倍	7.87	1.56

从表 4.1 可以看出，双倍加强对于该数据集是个比较合适的选择。三倍加强无论均值还是标准差性能的下降也从一个侧面反映了频带间整体区分度的差异并不至于特别的大。

4.4.2.2 整体区分度曲线

将双倍加强时得到的一组等错误率的值，代入公式（4-1）、（4-2）和（4-3）中，即可得到各个频带的区分度数值。由于频带的强调和弱化与相应三角滤波器设置的密集和稀疏正相关，因此以频带的整体区分度的倒数作为三角滤波器宽度的设置依据，可以得到当前性能驱动准则下在频率域各频带的分辨率（以三角滤波器宽度表示），如图 4.3 所示。

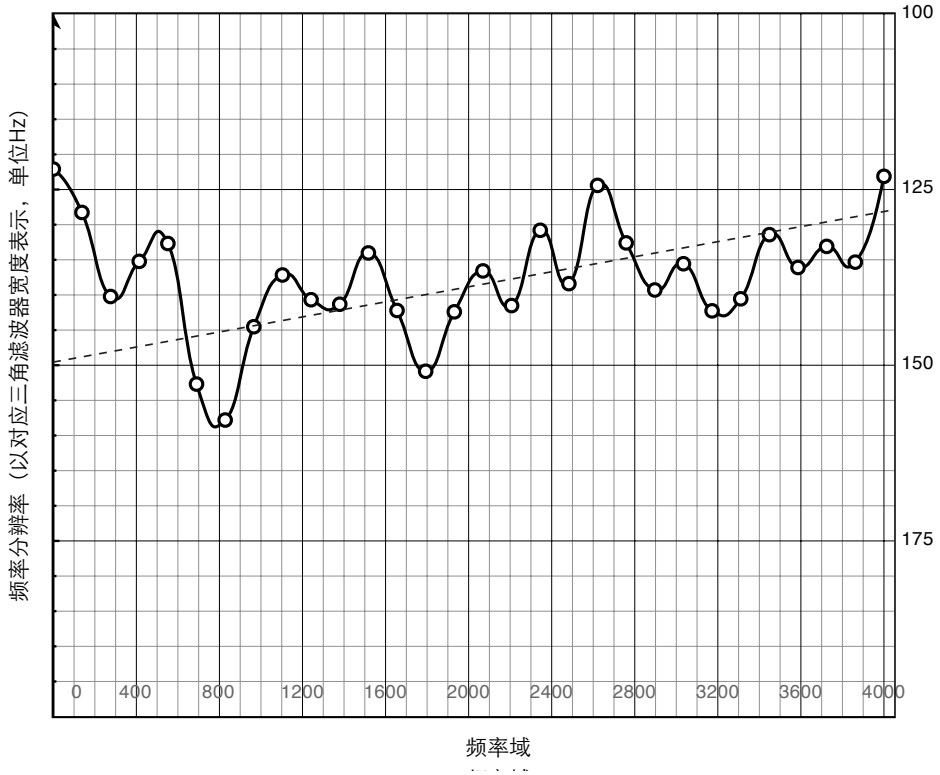

图4.3 性能驱动准则下的频带分辨率（Hz）

图 4.3 中虚线为各频带的分辨率数值拟合线。整体来看，与基于频带能量的 F-ratio 准则一致，高频区域的区分度总体依然高于低频区域。但很明显的是，这种差距远没有前者数值差异那么大。

对比图 3.8，可以看出来性能驱动准则下的三角滤波器宽度设置，在低频区域的变化类似于基于频带能量和 F-ratio 准则下的设置，尤其在频带 4 和频带 13 之间（400~1,600Hz 频率区域内），波峰波谷出现在的频率位置比较相近；而在高频区域则更类似于一种线性刻度的设置，分辨率区线在 130Hz 上下随机抖动，差异不明显。

4.4.3 实验结果

基于图 4.3 的三角滤波器设置进行频率弯折，提取倒谱特征（16 维倒谱加一阶差分）进行了说话人确认的实验，记为 Warping_Pfm_drvn，并与基线系统的 MFCC 特征，以及第 3 章中提出的 Warping_F_ratio 作了比较。在缺少足够的生理解剖学方面先验知识的前提下，比较这两种区分度准则的优劣，只有依靠 4.3 节所定义的时变说话人确认系统的性能指标。这三种倒谱系数特征在时变说话人确认系统中的性能对比见表 4.2。

表4.2 三种倒谱系数特征方法的比较（%）

倒谱特征方法	等错误率		下降率
	均值	标准差	均值*标准差
MFCC	8.80	1.86	--
Warping_F_ratio	7.77	1.54	26.90
Warping_Pfm_drvn	7.32	1.51	32.47

从表 4.2 中可以看出，强调低频弱化高频的 MFCC 特征性能表现最差，整体而言更强调高频区域的两种频率弯折方法均优于 MFCC。而如之前分辨率曲线图所分析，Warping_Pfm_drvn 在低频区域类似于 Warping_F_ratio，而之后又近似一种线性刻度变换。也许正是因为它将对高频的强调和对低频的弱化维持在了一个更加合适的程度，既体现出了局部的高低变化，总体趋势又比较平稳，使得在频率弯折的前提下，Warping_Pfm_drvn 的性能稍微好于 Warping_F_ratio。

4.5 小结

本章从基于频带能量和 F-ratio 准则的局限性出发，提出利用性能驱动的方式来确定频带整体区分度的可行性，并且探讨了针对频率弯折方式的性能驱动准则：单独加强某一频带构建说话人识别系统，并将系统性能作为该频带的整体区分度。

之后针对时变说话人识别任务的特点，分别从说话人确认任务和说话人辨认任务两个角度，定义了系统的整体性能评价指标。据此以说话人确认任务为例，确定了性能驱动准则下频带的整体区分度。本章最后对基线 MFCC 和两种准则下的频率弯折方法进行了性能对比。

第5章 基于区分性训练的滤波器输出加权方法的特征提取算法

5.1 引论

第 4 章中针对频率弯折方式采取的性能驱动准则，由于频带加强与否及加强程度是直接利用说话人识别系统的识别结果来确定的，故通过性能驱动准则选出的频率弯折方式通常能达到目的。但是这种性能驱动准则本质上相当于对所加强频带和加强幅度进行无先验知识的盲目搜索，且每次搜索中都需要经历提取特征、建模、打分等系统建立评估的全过程。假设所有 K 个频带都可以被同时加强或削弱，每个频带可有 P 种不同的加强、削弱的程度，那么通过盲目搜索找出该设置下的最优参数需要将全过程重复 K^P 次，几乎是无法求解的；即使将各个参数所有取值的排列组合视为离散变量，使用遗传算法（Mitchell, 1999）、模拟退火（龚光鲁等, 2007）等随机优化算法求解也仍然要面临着巨大的计算量，第 4 章中的做法也是在考虑这些因素后进行的一种简化。进一步地，由于这种性能驱动的准则完全依赖系统在开发集上的性能进行参数选择，不可避免的要遇到参数的过拟合（Over-fitting）问题，该性能驱动准则缺少能够控制系统参数推广能力（Generalization ability)的参数，只能通过使用 k 折交叉验证（k-fold cross validation）等方法来达成这一目的（Bishop, 2007），而这也会大大增加实验的计算量。

因此，我们需要一种更为系统的方式来进行特征层面参数的优化。这种方式既要与系统的性能直接相关（依然是性能驱动准则的模式），又要具有适当的计算复杂度，并且容易控制。语音识别领域的区分性特征提取算法 （DFE, Discriminative Feature Extraction）可以有效的达到这一目的（Biem and Katagiri, 1993）。

DFE 算法的基本思路与第 4 章中提出的性能驱动准则相似，即将识别结果反馈到特征提取模块，利用反馈进行参数优化。不同的是 DFE 算法可以利用经典的最小分类错误 MCE 区分性训练技术（Juang *et al*., 1997）。它将识别器输出的对数似然分嵌入光滑可导的 Sigmoid 函数中，使目标函数的形式接近于对识别错误进行记数的 0-1 分类错误函数，并同时保证目标函数对模型的参数（高斯混合成分的均值、方差等）光滑可导。于是利用基于梯度的推广的概率下降算法 （GPD, Generalized Probability Descent）进行优化，通过迭代寻找使目标函数对待优化参数的一阶导数为 0 的极值点来求得使目标函数收敛到局部最优解的参数值

（Snyman, 2005）。DFE 算法将特征层面的参数（如 MFCC 计算中 FFT 后各点的权值、高斯型滤波器的均值、方差、权重等）也视作模型的参数之一，即作为模型输入特征向量的参数嵌入优化的目标函数，并利用与 MCE 训练相同的思路进行优化。这样可以通过基于梯度的优化技巧高效的获得使系统错误率最小的特征参数。实验表明使用 DFE 算法求得使目标函数收敛的局部最优解通常只需要对训练数据进行 10 次左右的迭代（Miyajima *et al*., 2001; Beim *et al*., 2001）。

于是，我们试图将 DFE 算法与基于频带区分度进行时变说话人研究的思路相结合，进行基于 GMM-UBM 模型的时变说话人识别研究，即，将与频带区分度有关的特征参数嵌入到系统识别结果反馈得到的连续可导的目标函数中，利用基于梯度的优化方法对它们进行迭代优化，最终求得使目标函数达到局部最优的特征参数。与传统 DFE 方法不同，针对时变说话人研究中多次录音会话的特点，在 MCE 准则的基础之上提出了最小化会话方差 MSV 的准则，该准则利用 MCE 的目标函数求得的每次录音会话上的错误率，求使得不同会话间错误率的方差最小化的参数。由于实用中需要兼顾系统的识别率以及时变的鲁棒性， 将第 4 章中探讨过的系统性能指标进行扩展，提出了将 MCE 准则的目标函数与 MSV 准则的目标函数相乘作为目标函数（幂次由平衡因子控制），记做 MCE*MSV 准则。使得 MCE*MSV 准则最小化的特征参数，就是既能够保持说话人间区分度，又能提高系统对时变现象鲁棒性的特征参数，也即各个频带加强或弱化的程度，即频带的区分度。

考虑到 DFE 算法的特点，本章采用了滤波器输出加权的方式，将滤波器输出的权重作为特征参数，进行基于 MCE*MSV 准则的性能优化。

5.2 说话人识别中的区分性特征提取算法

5.2.1 区分性训练准则概述

在基于统计模型的机器学习和模式识别问题中，统计模型的设计，模型的训练准则和参数优化方法一直是模型模块所关注的核心内容（Bishop, 2007）。对于说话人识别中常用的 GMM 模型来说，它隐含假设特定说话人的语音帧作为独立同分布的随机过程符合 GMM 所描述的分布。这是由于缺少对说话人语音帧分布的先验知识，即真实分布的形式未知，所以利用不断增多的高斯分布的线性组合来在任意程度近似任何分布（Sorenson and Alspach, 1971; Juang *et al*., 1986），这样就引入了模型层面的第一种近似，即分布类型上的近似。使用 EM 算法基于最大似然准则 （ML, Maximum Likelihood）进行参数估计时，使用越多的训练样本

得到的模型可以越复杂，参数越可靠，故使用 GMM 近似真实分布的准确程度也受到训练样本数目的限制，有限的训练数据导致了第二种近似。第三，在使用经典的 EM 算法对指定规模的 GMM 模型的参数进行迭代优化时，由于真实语音数据的复杂性以及 EM 算法本身的特点（总是寻找导数为 0 的参数），算法很容易收敛到局部最优点（Bishop, 2007），故此时存在局部最优参数到全局最优参数的近似。综上所述，在使用 ML 准则和 EM 算法训练 GMM 模型参数时常常具有较高的错误率。

为使系统具有更好的性能，常通过改进前述近似从模型角度降低系统错误率。对于已有的系统，常修改模型的训练准则及其相应的优化算法。最常见的训练准则包括区分性训练准则和大边距准则（LM, Large Marginal）。常见的区分性准则包括：最大互信息准则 MMIE（Maximum Mutual Information Estimation）（Valtchev *et al*., 1997），通过最大化语音帧 $\vec{x}$ 对其所属分类 $\vec{y}$ 的后验概率 $P(\vec{y}|\vec{x})$ 来提高模型的区分度，使用扩展的 EM 算法进行训练；MCE 准则，利用识别结果作为反馈来修改模型参数，并使用 GPD 算法进行优化；最小化音子错误准则 MPE（Minimum Phone Error）（Povey and Woodland, 2002），利用动态规划对齐获得语音识别结果的音子错误率对后验概率进行加权并通过扩展 EM 算法来优化语音识别中的声学模型；最新的区分性训练准则还包括非均一化 MCE 准则（Fu *et al*., 2012）和软边距准则（Li and Lee, 2007）等。近年来，由于 SVM 的成功，基于严格的统计学习理论（Vapnik, 1998），在结构风险最小化意义下具有最优推广性的大边距准则也在其它模型中获得了广泛的应用，包括大边距的 GMM、HMM（Sha, 2006）等。在以上准则中，MCE 准则与经典的 MMIE 等准则相比性能更好（Schluuter, 2001）；与大边距准则等新准则相比，MCE 计算简单，训练效率更高；同时目标函数形式较为简单，易于改进优化，故本章基于 MCE 准则的框架来进行特征参数的优化。本节其余部分以基于 GMM 模型的说话人识别系统为例，详细介绍了 MCE 区分性训练准则、优化算法以及该准则在 DFE 算法中的应用和对特征参数的优化策略。

5.2.2 MCE区分性训练算法

在 MCE 训练中，记一段语音的输入观测特征向量为 X，Λ 为 GMM 模型的参数，记 $g_i(X;\Lambda)$ 为 X 在第 i 个说话人模型上得到的对数似然度得分，其中 $i=1,2,\ldots,I$，I 为说话人总数目。定义 X 的误分类函数为：

$$d_i(X) = -g_i(X;\Lambda) + \ln\left[\frac{1}{I-1}\sum_{j,j\neq i}\exp[g_j(X;\Lambda)\eta]\right]^{1/\eta}. \qquad (5\text{-}1)$$

其中，η 是任取的正数，η 增大时 $d_i(X)$ 减小，可以用于对误分类函数进行控

制。另外由极限知识易知，当$\eta \to \infty$时有：$d_i(X) \to -g_i(X;\Lambda) + \max_{j,j\neq i} g_j(X;\Lambda)$。即极限为目标说话人模型和最有可能的假冒说话人模型之间的对数似然分之差；而当η减小时则可以在$d_i(X)$的计算中考虑特征向量在更多的假冒说话人模型上的得分。易知$d_i(X) \le 0$时说明模型对特征向量被正确分类，否则说明发生了分类错误。在说话人识别应用中，可以通过调整η和I来改变计算误分类函数时所需考虑的模型数，从而在准则层面对不同的假冒说话人产生的影响进行平衡。

MCE 准则通过将公式（5-1）嵌入到如下的 Sigmoid 函数中来定义目标函数：

$$l_i(X;\Lambda) = l(d_i(X)) = \frac{1}{1+\exp(-\alpha \cdot d_i(X))}. \qquad (5\text{-}2)$$

上式中，$\alpha > 0$，是可以用来控制$l_i(X;\Lambda)$函数形状的参数。若α很小（比如 0.01），则$l_i(X;\Lambda)$在$d_i(X)$的主要值域上都近似于一个较平缓的线性函数，此时对数值不同的$d_i(X)$，得到的$l_i(X;\Lambda)$有较明显的差异；而当α较大时（比如 5.0），$l_i(X;\Lambda)$近似于一个 0-1 错误函数，即正确识别的程度越高（$d_i(X)$越小），则$l_i(X;\Lambda)$越接近于 0，否则随着错误程度的上升$l_i(X;\Lambda)$逐渐趋近于 1，相当于起到对分类错误进行累加计数的效果，如图 5.1 所示。经验表明α较小时得到的模型参数在未知数据上可能具有更好的推广性（McDermott and Hazen, 2004）。

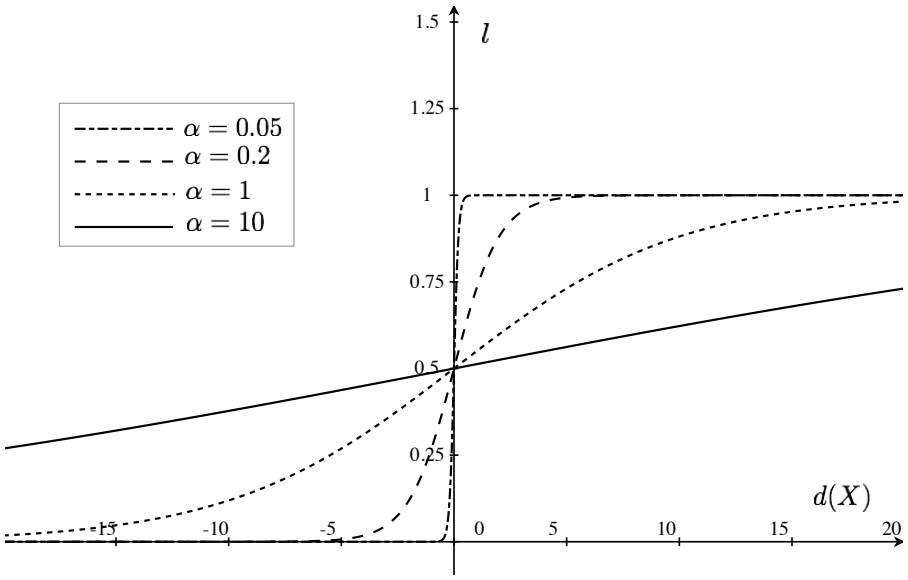

图5.1　目标函数随α变化示意图

自然的，可以利用（5-2）将一段语音的特征向量是否发生误分类记为：

$$l(X;\Lambda) = \sum_{i=1}^{I} l_i(X;\Lambda) \cdot 1(X \in \text{speaker}_i). \qquad (5\text{-}3)$$

其中1(·)是示性函数，当其括号内条件为真时值为 1，否则为 0。进一步的，可以定义 MCE 训练集上的目标函数，记为：

$$L(\Lambda)=\frac{1}{U}\sum_{u=1}^{U}l(X_u;\Lambda). \tag{5-4}$$

这是 MCE 训练集上的错误率，称为经验错误率（Vapnik, 1998）。需要说明的是，这儿的训练集指的是进行 MCE 区分性训练的数据集合，而在通常的实验中称这部分数据为开发集数据，用以得到中间参数。

目标函数（5-4）近似于 0-1 错误函数组成的错误率，意义明确。且若假设训练集中所有特征向量 X 等概率出现，则当 $U\to\infty$ 时可以得到：

$$\begin{aligned}L(\Lambda)&=\frac{1}{U}\sum_{u=1}^{U}l(X_u;\Lambda)\\&\to\int l(X;\Lambda)dP(X)=E_X\{l(X;\Lambda)\}=\hat{L}(\Lambda).\end{aligned} \tag{5-5}$$

即当训练集无穷大时，经验风险收敛到近似 0-1 函数的数学期望，于是 $\hat{L}(\Lambda)$ 也称作期望风险（Juang *et al*., 1997; Vapnik, 1998），是在训练和使用中理论上得到的错误率。

于是在 MCE 准则中，可以通过控制 α 来容易地控制模型的推广能力，推广能力越好的模型在未知数据上的性能也就越好（Bishop, 2007）。同时 MCE 准则的目标函数光滑可导，可以利用基于梯度的优化算法，通过对目标函数求导以迭代获得使模型性能较好的局部最优参数。此外，MCE 准则的形式可以使其满足基于梯度的 GPD 优化方法的所有应用条件，故可以使用 GPD 进一步加速优化过程。

5.2.3 基于梯度的GPD模型参数优化算法

GPD 又称作统计近似，是一种无约束的随即优化算法。算法用于寻找随机变量（$l(X;\Lambda)$）的数学期望函数（$\hat{L}(\Lambda)$）取得最值时的根。算法假设参数满足以下条件（Robbins and Monro, 1951; Chou *et al*., 1992; Juang *et al*., 1997）：

（1）$\sum_{t=1}^{\infty}\varepsilon_t=\infty$，$\sum_{t=1}^{\infty}\varepsilon_t^2<\infty$ 且 $\varepsilon_t>0$；

（2）存在常数 V，满足 $0\le V<\infty$，使得对所有 t 都满足：

$$R_t(\varepsilon_t,\theta_t)=\left\langle\nabla l(X;\Lambda_n),H(X;\Lambda_n+\varepsilon_n\theta_n\nabla l(X;\Lambda_n))\nabla l(X;\Lambda_n)\right\rangle\le V.$$

其中 H 是由二阶偏导数组成的海森矩阵，而 $R_t(\varepsilon_t,\theta_t)$ 是两个向量的夹角。

（3）存在唯一的 Λ^* 使得 $\Lambda^*=\arg\min_{\Lambda}E_X\{l(X;\Lambda)\}$，即唯一存在使模型分类错

误最小的参数。此时由连续函数最小值的性质有 $\nabla L(\Lambda)|_{\Lambda=\Lambda^*}=\nabla E_X\{l(X;\Lambda)\}|_{\Lambda=\Lambda^*}=0$。

定义模型序列 Λ_t 为：

$$\Lambda_{t+1}=\Lambda_t-\varepsilon_t U_t\nabla l(X_t,\Lambda)|_{\Lambda=\Lambda_t}. \tag{5-6}$$

Λ_t 以概率 1 收敛为 Λ^*，U_t 为正定矩阵。实用中常取 $\varepsilon_t=1/t$，U_t 为单位矩阵，在每次遍历完训练集数据之后都要对训练集所有数据进行随机重排序，以增强 $l(X;\Lambda)$ 的随机性。

对于 GMM 模型，参数 Λ 包括模型中每个高斯混合的均值、方差和权值。记 l 个高斯混合的均值和权值分别为 μ_l 和 w_l，其协方差矩阵第 j 行第 k 列的元素为 $\delta_{l,j,k}^2$。这些参数并不是无约束的变量，它们需要满足以下两个条件：

（1）$\sum_l w_l=1$；

（2）标准差 $\delta_{l,j,k}>0$。

一般可通过可逆变换将有约束条件的参数变换到等价的无约束参数空间中的来解决（Juang *et al*., 1997; Miyajima *et al*., 2001）。记变换空间的参数分别为 $\tilde{\mu}_l$、$\tilde{w}_l$ 和 $\tilde{\delta}_{l,j,k}$。则，对于均值、标准差及权值有：

$$\mu_l\to\tilde{\mu}_l,\ \mu_l=\tilde{\mu}_l\delta_l. \tag{5-7}$$

$$\delta_{l,j,k}\to\tilde{\delta}_{l,j,k},\ \delta_{l,j,k}=e^{\tilde{\delta}_{l,j,k}}. \tag{5-8}$$

$$w_l\to\tilde{w}_l,\ w_l=\frac{e^{\tilde{w}_l}}{\sum_r e^{\tilde{w}_r}}. \tag{5-9}$$

这样，对训练集上的某特征向量 X_n，求 $l(X_n;\tilde{\Lambda})$ 对上述变换后参数 $\tilde{\Lambda}_t$ 的梯度，利用（5-6）式得到更新的参数 $\tilde{\Lambda}_{t+1}$，再从（5-7）～（5-9）式得到逆运算将 $\tilde{\Lambda}_{t+1}$ 变回有约束的原参数空间即可。当对训练集中的每一特征向量都重复过上述步骤后，即完成一次迭代。下次迭代开始前先随机打乱所有数据的顺序，再依次对每一特征向量重复以上过程，直到模型参数收敛，就可以完成 MCE 区分性训练。

5.2.4 DFE算法

如引论所述，基于识别系统性能的特征参数调整方法与系统性能关系更为直接，常常具有更好的效果。但应用这种方法时，对每组特征参数的评估都要重新经过系统建模、打分的过程，运算开销较大。当模型参数较复杂时，通过简单重复实验来得到最优参数的计算开销很大，削弱了方法的实用价值。为了解决这一

问题，Miyajima 等人（2001）基于 MCE 区分性训练的技术提出了 DFE 算法。

DFE 算法优化的目标参数是 FFT 后每个点的权重、滤波器、频带等特征层次的参数，下文通称之为特征参数，记为C。特征参数作用于较初级的特征以得到最终的特征X。将较初级的特征，如 FFT 后每点的输出、滤波器的输出等称为初级特征，记为Y。则有$X = f(Y;C)$，f是初级特征在特征参数的作用下得到最终特征的函数关系。则 DFE 算法的主要思想是将C通过X嵌入 MCE 的目标函数中，通过与优化模型参数Λ相似的技术来求得能够使分类错误率最小的特征参数C^*。更形式化的，将$X = f(Y;C)$代入公式（5-1）～（5-4），即得到 DFE 算法的目标函数和优化方法如下：

$$d_i(f(Y;C)) = -g_i(f(Y;C);\Lambda) + \ln\left[\frac{1}{I-1}\sum_{j,j\neq i}\exp[g_j(f(Y;C);\Lambda)\eta]\right]^{1/\eta}. \quad (5\text{-}10)$$

$$l_i(f(Y;C);\Lambda) = l(d_i(f(Y;C))) = \frac{1}{1+\exp(-\alpha\cdot d_i(f(Y;C)))}. \quad (5\text{-}11)$$

$$l(f(Y;C);\Lambda) = \sum_{i=1}^{I} l_i(f(Y;C);\Lambda)\cdot 1(X = f(Y;C)\in \mathrm{spk}_i). \quad (5\text{-}12)$$

$$L(C;\Lambda) = \frac{1}{U}\sum_{u=1}^{U} l(f(Y_u;C);\Lambda). \quad (5\text{-}13)$$

应用 GPD 算法进行特征参数优化的公式变为：

$$C_{t+1} = C_t - \varepsilon_t U_t \nabla l(f(Y_t;C),\Lambda)|_{C=C_t} \quad (5\text{-}14)$$

这样从梯度的反方向进行参数搜索，就可以高效的求出使得系统错误率最低的特征参数（Snyman, 2005）。序列C_t最终将收敛到使得错误率最小化的特征参数C^*。

进一步的，还可以将特征参数和模型参数同时作为 MCE 区分性训练的优化目标参数，即同时优化C和Λ，以削弱说话人识别系统中特征提取模块和模型模块相互独立的假设。虽然这种假设可以使系统各个模块的设计相互独立，从而降低设计难度，但研究表明模块化设计下取得的系统参数很难获得系统全局最优性能（Biem *et al*., 2001）。在系统设计完成后，使用 MCE 准则，利用 DFE 和通常的区分性训练技术对特征参数和模型参数联合进行联合优化，可以有效削弱各模块间相互独立的假设，从而提高系统性能（Biem *et al*., 2001）。

受上述 MCE-DFE 算法的启发，针对时变说话人识别任务的特点，扩展这种区分性训练的思路，利用系统误识率（性能）作为反馈直接寻找既能够突出说话人个性信息又能够对时变现象鲁棒的特征参数，详述如下。

5.3 时变说话人识别的区分性准则

5.3.1 最小会话方差准则

从 5.2 节的分析可以看出，MCE-DFE 算法可以有效寻找使得模型在训练数据上总错误率最低的参数。在时变说话人识别问题中，假设共有说话人的 S 次录音会话数据，系统在这些数据上的识别错误率为 $L(C;\Lambda)$，记系统在第 s 次会话数据上的错误率为 $L^{(s)}(C;\Lambda)$，即：

$$L^{(s)}(C;\Lambda)=\frac{1}{U^{(s)}}\sum_{u=1}^{U^{(s)}}l(f(Y_u^{(s)};C);\Lambda). \tag{5-15}$$

其中 $U^{(s)}$ 和 $Y_u^{(s)}$ 分别是第 s 次会话中的所有语音数据总数和其中第 u 句会话的初始特征。则有：

$$L(C;\Lambda)=\frac{1}{S}\sum_{s=1}^{S}L^{(s)}(C;\Lambda). \tag{5-16}$$

假设使用基于 MCE 准则的 DFE 技术得到的特征参数为 C'，有 $L(C;\Lambda)\geq L(C';\Lambda)$。但对于某次录音会话数据上的分类错误 $L^{(s)}$，可能有 $L^{(s)}(C;\Lambda)>L^{(s)}(C';\Lambda)$，$L^{(s)}(C;\Lambda)=L^{(s)}(C';\Lambda)$ 或 $L^{(s)}(C;\Lambda)<L^{(s)}(C';\Lambda)$。也就是说 L 在达到最小化的同时，系统在不同录音会话数据上的性能差异可能会有所增加。也就是说，系统在总错误率下降的同时对时变现象的鲁棒性有可能会变差。类似的情形也会发生在基于 MCE 准则对模型参数进行区分性训练的情况下。

为了解决这一问题，保证算法能够找到对时变现象较为鲁棒的参数，即能够获得使各次录音会话间错误率较为接近的参数，提出了最小化会话错误率方差的准则 MSV。其目标函数 $V(C;\Lambda)$ 如下：

$$\begin{aligned}V(C;\Lambda)&=\frac{1}{S}\sum_{s=1}^{S}\left[L^{(s)}(C;\Lambda)-L(C;\Lambda)\right]^2\\&=\frac{1}{S}\sum_{s=1}^{S}\left[L^{(s)}(C;\Lambda)-\frac{1}{S}\sum_{n=1}^{S}L^{(n)}(C;\Lambda)\right]^2.\end{aligned} \tag{5-17}$$

即 MSV 准则优化的目标函数是各次会话间错误率的方差。通过最小化错误率的方差 $V(C;\Lambda)$ 就可以达到使参数在各次会话间错误率尽可能接近的目的。

5.3.2 MCE*MSV准则

考虑到进行参数优化的目标不仅仅是寻找对时变现象足够鲁棒的参数，也要保证特征参数本身能够突出说话人个性信息，时变说话人识别任务终究还是说话人识别任务，务必使得说话人识别系统具有较好的识别率。如果仅仅考虑从方差角度选取使各次会话间错误率尽可能接近的参数，那么极端情况下可能选出在各次会话中错误率均为100%的参数，完全没有应用价值。所以，特征的选择标准，即优化的目标函数应该既能够在识别率和时变鲁棒性之间取得平衡。第 3 章的性能驱动准则也证实了等错误率乘以各次录音会话间错误率的标准差是评估特征参数的一个较好标准。受此启发，在 MCE 和 MSV 准则的基础上提出了 MCE*MSV 准则作为时变说话人识别任务的参数选择标准。准则定义如下：

$$\begin{aligned} LV(C;\Lambda) &= L(C;\Lambda)\cdot V(C;\Lambda)^{\beta} \\ &= L(C;\Lambda)\cdot\left\{\frac{1}{S}\sum_{s=1}^{S}\left[L^{(s)}(C;\Lambda)-\frac{1}{S}\sum_{n=1}^{S}L^{(n)}(C;\Lambda)\right]^2\right\}^{\beta} \\ &= \frac{1}{S^{\beta+1}}\left[\sum_{n=1}^{S}L^{(n)}(C;\Lambda)\right]\cdot\left\{\sum_{s=1}^{S}\left[L^{(s)}(C;\Lambda)-\frac{1}{S}\sum_{n=1}^{S}L^{(n)}(C;\Lambda)\right]^2\right\}^{\beta}. \end{aligned} \tag{5-18}$$

其中，β是对 MCE 和 MSV 准则的影响进行平衡的平衡因子，$0<\beta<1$；β取 0.5 时即为第 3 章中使用的评价指标——均值乘以标准差。

5.4 MCE*MSV准则下的参数训练方法

5.4.1 目标函数的偏导数

从（5-6）和（5-14）式易知，使用基于梯度的参数优化方法时需要计算目标函数对各个待优化参数的导数。于是，对（5-18）式的特征参数和模型参数分别求导数。为方便起见，记$\Theta=C\cup\Lambda$为特征参数和模型参数的并集，即所有待优化的参数集合；简记$l_i(f(Y_u^{(s)};C);\Lambda)$为$l_i$，$d_i(f(Y_u^{(s)};C);\Lambda)$为$d_i$，$g_i(f(Y_u^{(s)};C);\Lambda)$为$g_i$。

首先，依据求导数的链式法则和乘法原理，有：

$$\begin{aligned}\frac{\partial(LV(C;\Lambda))}{\partial\Theta} &= L(C;\Lambda)\frac{\partial V(C;\Lambda)^{\beta}}{\partial\Theta}+V(C;\Lambda)^{\beta}\frac{\partial L(C;\Lambda)}{\partial\Theta}\\ &= L(C;\Lambda)V(C;\Lambda)^{\beta-1}\frac{\partial V(C;\Lambda)}{\partial\Theta}+V(C;\Lambda)^{\beta}\frac{\partial L(C;\Lambda)}{\partial\Theta}.\end{aligned} \tag{5-19}$$

即计算 MCE*MSV 准则目标函数对各参数的偏导数，需要分别计算 MCE 准则目标函数和 MSV 准则目标函数对这些参数的偏导数。

对于 MSV 准则的目标函数$V(C;\Lambda)$，有：

$$\begin{aligned}\frac{\partial V(C;\Lambda)}{\partial\Theta} &= \frac{\partial\left\{\frac{1}{S}\sum_{s=1}^{S}\left[L^{(s)}(C;\Lambda)-\frac{1}{S}\sum_{n=1}^{S}L^{(n)}(C;\Lambda)\right]^2\right\}}{\partial\Theta}\\ &= \frac{2}{S}\sum_{s=1}^{S}\left[L^{(s)}(C;\Lambda)-\frac{1}{S}\sum_{n=1}^{S}L^{(n)}(C;\Lambda)\right]\\ &\quad\bullet\left[\frac{\partial L^{(s)}(C;\Lambda)}{\partial\Theta}-\frac{1}{S}\sum_{n=1}^{S}\frac{\partial L^{(n)}(C;\Lambda)}{\partial\Theta}\right].\end{aligned} \tag{5-20}$$

从（5-20）式可以看出$V(C;\Lambda)$的求导依然要归结到$L^{(s)}(C;\Lambda)$的求导上。而事实上$L(C;\Lambda)$的求导公式：

$$\frac{\partial L(C;\Lambda)}{\partial\Theta}=\frac{\partial\left\{\frac{1}{S}\sum_{s=1}^{S}L^{(s)}(C;\Lambda)\right\}}{\partial\Theta}=\frac{1}{S}\sum_{s=1}^{S}\frac{\partial L^{(s)}(C;\Lambda)}{\partial\Theta}. \tag{5-21}$$

可见，最终目标函数的求导都归结于$L^{(s)}(C;\Lambda)$的求导上来。继续推导$L^{(s)}(C;\Lambda)$的偏导数，可得：

$$\begin{aligned}\frac{\partial L^{(s)}(C;\Lambda)}{\partial\Theta} &= \frac{\partial\left[\frac{1}{U^{(s)}}\sum_{u=1}^{U^{(s)}}l(f(Y_u^{(s)};C);\Lambda)\right]}{\partial\Theta}\\ &= \frac{1}{U^{(s)}}\sum_{u=1}^{U^{(s)}}\frac{\partial l(f(Y_u^{(s)};C);\Lambda)}{\partial\Theta}.\end{aligned} \tag{5-22}$$

同时，有：

$$\frac{\partial l(f(Y_u^{(s)};C);\Lambda)}{\partial\Theta}=\frac{\partial l_i}{\partial\Theta}=\frac{\partial l_i}{\partial d_i}\frac{\partial d_i}{\partial\Theta}=\alpha\cdot l_i(1-l_i)\frac{\partial d_i}{\partial\Theta}. \tag{5-23}$$

而 d_i 的求导公式推导如下：

$$\begin{aligned}\frac{\partial d_i}{\partial \Theta} &= \frac{\partial\left\{-g_i+\ln\left[\frac{1}{I-1}\sum_{j,j\neq i}\exp(g_j\eta)\right]^{1/\eta}\right\}}{\partial\Theta}\\ &= -\frac{\partial g_i}{\partial\Theta}+\frac{1}{\eta}\frac{\partial\ln\left[\frac{1}{I-1}\sum_{j,j\neq i}\exp(g_j\eta)\right]}{\partial\Theta}\\ &= -\frac{\partial g_i}{\partial\Theta}+\frac{1}{\eta}\frac{1}{\sum_{j,j\neq i}\exp(g_j\eta)}\frac{\sum_{j,j\neq i}\partial\exp(g_j\eta)}{\partial\Theta}\\ &= -\frac{\partial g_i}{\partial\Theta}+\frac{1}{\eta}\frac{\eta}{\sum_{j,j\neq i}\exp(g_j\eta)}\left[\sum_{j,j\neq i}\exp(g_j\eta)\frac{\partial g_j}{\partial\Theta}\right]\\ &= -\frac{\partial g_i}{\partial\Theta}+\sum_{j,j\neq i}\left[\frac{\exp(g_j\eta)}{\sum_{k,k\neq i}\exp(g_k\eta)}\frac{\partial g_j}{\partial\Theta}\right].\end{aligned} \tag{5-24}$$

综合公式（5-19）～（5-24）可以发现，计算出 g_i 的偏导数就可算出最终目标函数的导数。

而对于说话人识别中常用的 GMM 模型来说，记 $Y_u^{(s)}=\{y_{u,1}^{(s)},y_{u,2}^{(s)},\ldots,y_{u,T_u^{(s)}}^{(s)}\}$，$T_u^{(s)}$ 为第 s 次录音会话中第 u 句话的总帧数。则 $g_i(f(Y_u^{(s)};C);\Lambda)$ 定义为：

$$g_i(f(Y_u^{(s)};C);\Lambda)=\frac{1}{T_u^{(s)}}\sum_{t=1}^{T_u^{(s)}}\ln b_i(f(y_{u,t}^{(s)};C)). \tag{5-25}$$

其中 $b_i(f(y_{ut}^{(s)};C))$ 为语音的第 t 帧在第 i 个说话人的 GMM 模型上的得分。记其第 l 个高斯混合为：

$$\begin{aligned}N_{i,l} &= N_{i,l}(\mu_{i,l},\Sigma_{i,l};f(y_{u,t}^{(s)};C))\\ &= \frac{\exp\left\{\frac{-\left[f(y_{u,t}^{(s)};C)-\mu_{i,l}\right]^T\Sigma_{i,l}^{-1}\left[f(y_{u,t}^{(s)};C)-\mu_{i,l}\right]}{2}\right\}}{\sqrt[d]{2\pi}\sqrt{|\Sigma_{i,l}|}}.\end{aligned} \tag{5-26}$$

则有：

$$b_i(f(y_{u,t}^{(s)};C))=\sum_{l=1}^{L}w_{i,l}N_{i,l}. \tag{5-27}$$

其中，D 是特征的维度，$w_{i,l}$ 是 GMM 模型中每个高斯混合的权值，有 $\sum_{l=1}^{L} w_{i,l} = 1$。

公式（5-27）即一条语音中各帧在各个高斯混合上的似然度的线性组合之和。由于本文主要研究的倒谱系数特征是各维间相互独立的，于是协方差矩阵 $\Sigma_{i,l}$ 可以写成 $\{(\delta_{i,l,d})^2\}|_{d=1}^{D}$，即协方差矩阵简化为主对角矩阵。并简记 $b_i(f(y_{u,t}^{(s)};C))$ 为 b_i。于是公式（5-27）可以简化为：

$$\begin{aligned} b_i &= \sum_{l=1}^{L} w_{i,l} N_{i,l} = \sum_{l=1}^{L} w_{i,l} N_{i,l}(\mu_{i,l,d}|_{d=1}^{D}, \{(\delta_{i,l,d})^2\}|_{d=1}^{D}; f(y_{u,t}^{(s)};C)) \\ &= \sum_{l=1}^{L} \frac{w_{i,l}}{\sqrt[D]{2\pi}\prod_{d=1}^{D}\delta_{i,l,d}} \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left[\frac{f(y_{u,t,d}^{(s)};C)-\mu_{i,l,d}}{\delta_{i,l,d}}\right]^2\right\}. \end{aligned} \tag{5-28}$$

于是，利用（5-26）、（5-27）和（5-28）式求 g_i 的偏导数，可得：

$$\begin{aligned} \frac{\partial g_i}{\partial \Theta} &= \frac{1}{T_u^{(s)}}\sum_{t=1}^{T_u^{(s)}} \frac{\partial \ln b_i}{\partial \Theta} = \frac{1}{T_u^{(s)}}\sum_{t=1}^{T_u^{(s)}} \frac{1}{b_i}\frac{\partial b_i}{\partial \Theta} \\ &= \frac{1}{T_u^{(s)}}\sum_{t=1}^{T_u^{(s)}} \frac{1}{b_i}\sum_{l=1}^{L} \frac{\partial w_{i,l} N_{i,l}(\mu_{i,l,d}|_{d=1}^{D}, \{(\delta_{i,l,d})^2\}|_{d=1}^{D}; f(y_{u,t}^{(s)};C))}{\partial \Theta}. \end{aligned} \tag{5-29}$$

5.4.2 模型参数的训练

当优化模型参数时，需要考虑（5-7）～（5-9）给出的参数变换。即用 $\tilde{\mu}$、$\tilde{\delta}$ 和 $\tilde{w}$ 在（5-29）中分别取代 μ、δ 和 w，再分别对 $\tilde{\mu}$、$\tilde{\delta}$ 和 $\tilde{w}$ 求偏导。令 Θ 分别为 $\tilde{\mu}$、$\tilde{\delta}$ 和 $\tilde{w}$，代换后求导可得：

$$\begin{aligned} \frac{\partial g_i}{\partial \tilde{\mu}_{i,l,d}} &= \frac{1}{T_u^{(s)}}\sum_{t=1}^{T_u^{(s)}} \frac{1}{b_i}\sum_{l=1}^{L} \frac{\partial w_{i,l} N_{i,l}(\delta_{i,l,d}\tilde{\mu}_{i,l,d}|_{d=1}^{D}, \{(\delta_{i,l,d})^2\}|_{d=1}^{D}; X_{u,t,d}^{(s)}|_{d=1}^{D})}{\partial \tilde{\mu}_{i,l,d}} \\ &= \frac{1}{T_u^{(s)} b_i}\sum_{t=1}^{T_u^{(s)}}\sum_{l=1}^{L} \frac{w_{i,l}\left[\frac{X_{u,t,d}^{(s)}-\mu_{i,l,d}}{\delta_{i,l,d}}\right]}{\sqrt[D]{2\pi}\prod_{d=1}^{D}\delta_{i,l,d}} \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left[\frac{X_{u,t,d}^{(s)}-\mu_{i,l,d}}{\delta_{i,l,d}}\right]^2\right\}. \end{aligned} \tag{5-30}$$

$$\begin{aligned} \frac{\partial g_i}{\partial \tilde{\delta}_{i,l,d}} &= \frac{1}{T_u^{(s)} b_i}\sum_{t=1}^{T_u^{(s)}}\sum_{l=1}^{L} \frac{\partial w_{i,l} N_{i,l}(\mu_{i,l,d}|_{d=1}^{D}, \{\exp\{2\tilde{\delta}_{i,l,d}\}\}|_{d=1}^{D}; X_{u,t,d}^{(s)}|_{d=1}^{D})}{\partial \tilde{\delta}_{i,l,d}} \\ &= \frac{1}{T_u^{(s)} b_i}\sum_{t=1}^{T_u^{(s)}}\sum_{l=1}^{L} \frac{w_{i,l}\left[\left(\frac{X_{u,t,d}^{(s)}-\mu_{i,l,d}}{\delta_{i,l,d}}\right)^2-1\right]}{\sqrt[D]{2\pi}\prod_{d=1}^{D}\delta_{i,l,d}} \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left[\frac{X_{u,t,d}^{(s)}-\mu_{i.l,d}}{\delta_{i,l,d}}\right]^2\right\}. \end{aligned} \tag{5-31}$$

$$\frac{\partial g_i}{\partial \tilde{w}_{i,l}} = \frac{\sum_{t=1}^{T_u^{(s)}}\sum_{l=1}^{L}\frac{\partial \frac{\exp\{\tilde{w}_{i,l}\}N_{i,l}(\mu_{i,l,d}|_{d=1}^{D},\{(\delta_{i,l,d})^2\}|_{d=1}^{D};X_{u,t,d}^{(s)}|_{d=1}^{D})}{\sum_{k=1}^{L}\exp\{\tilde{w}_{i,k}\}}}{\partial \tilde{w}_{i,l}}}{T_u^{(s)}b_i}$$

$$= \frac{\sum_{t=1}^{T_u^{(s)}}\exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left[\frac{X_{u,t,d}^{(s)}-\mu_{i,l,d}}{\delta_{i,l,d}}\right]^2\right\}\frac{\exp\{\tilde{w}_{i,l}\}}{\sum_{k=1}^{L}\exp\{\tilde{w}_{i,k}\}}}{T_u^{(s)}b_i\sqrt[D]{2\pi}\prod_{d=1}^{D}\delta_{i,l,d}} \tag{5-32}$$

$$-\frac{1}{T_u^{(s)}b_i}\sum_{t=1}^{T_u^{(s)}}\sum_{j=1}^{L}\frac{\exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left[\frac{X_{u,t,d}^{(s)}-\mu_{i,l,d}}{\delta_{i,l,d}}\right]^2\right\}\left(\frac{\exp\{\tilde{w}_{i,l}\}}{\sum_{k=1}^{L}\exp\{\tilde{w}_{i,k}\}}\right)^2}{\sqrt[D]{2\pi}\prod_{d=1}^{D}\delta_{i,l,d}}.$$

当利用基于梯度的优化算法对模型参数进行一次迭代更新后，用公式（5-7）～（5-9）的逆变换将 $\tilde{\mu}$、$\tilde{\delta}$ 和 $\tilde{w}$ 代换回 μ、δ 和 w。即完成了优化算法的一次迭代。这些逆变换的公式为：

$$\tilde{\mu}_{i,l,d} = \frac{\mu_{i,l,d}}{\delta_{i,l,d}}. \tag{5-33}$$

$$\tilde{\delta}_{i,l,d} = \ln \delta_{i,l,d}. \tag{5-34}$$

$$\tilde{w}_{i,l} = \ln w_{i,l}. \tag{5-35}$$

这样就可以完成基于 MCE*MSV 准则对 GMM 模型参数的优化。

5.4.3 特征参数的训练

对于 DFE 方法中的特征参数，通常有多种选择方法：Miyajima 等人（2001）提出对 FFT 后每个频率点分别进行加权，将权值向量作为特征参数；还提出了使用高斯滤波器取代常用的三角滤波器，将每个高斯滤波器的均值、方差和对滤波器输出进行加权的权重作为特征参数。对于将 FFT 输出频率点的权值作为特征参数的方法，研究表明这种方法很容易产生较严重的参数过拟合现象，因而性能提高有限（Biem *et al*., 2001）；而将高斯滤波器的参数作为特征参数的方法则由于滤波器参数较多且有多种类型，所以通常实现比较复杂，且与经典的 MFCC 等使用的三角滤波器不同，使得方法既不容易广泛应用到已有系统中，又不容易与经典的基于三角滤波器的方法相比较。而其它特征优化方法中常见的对基于三角滤波器进行频率弯折的方法又由于三角滤波器函数连续但不可导，故不能使用 DFE 算

法对弯折频谱进行优化。

综上，本章选择对数运算后三角滤波器输出的权重作为特征参数。相比于对 FFT 的所有频率点分别进行加权的方法，滤波器对频带数目进行了压缩，参数数目减少，容易避免参数过拟合。

设三角滤波器的数目为 H，则初级特征为 H 个三角滤波器的输出进行对数运算后所构成的 H 维向量 $Y_{u,t}^{(s)}=\{y_{u,t,1}^{(s)},y_{u,t,2}^{(s)},\ldots,y_{u,t,H}^{(s)}\}$。此时初级特征 $Y_{u,t}^{(s)}$ 到特征向量 $X_{u,t}^{(s)}$ 的函数映射即为离散余弦变换，如图 1.3 所示。对 $Y_{u,t}^{(s)}$ 进行加权的权值向量 $C=\{c_1,c_2,\ldots,c_H\}$即为特征参数向量。由于三角滤波器的输出的权重可以是任意实数，所以没有必要对权重进行参数变换。于是上述 $Y_{u,t}^{(s)}$ 到 $X_{u,t}^{(s)}$ 的函数关系可表示为：

$$\begin{aligned}X_{u,t}^{(s)}&=\{x_{u,t,1}^{(s)},x_{u,t,2}^{(s)},\ldots,x_{u,t,D}^{(s)}\}=f(Y_{u,t}^{(s)};C)\\&=\left(\sum_{h=1}^{H}y_{u,t,h}^{(s)}c_h\cos\left(\frac{\pi(h+0.5)}{H}\right),\cdots,\sum_{h=1}^{H}y_{u,t,h}^{(s)}c_h\cos\left(\frac{D\pi(h+0.5)}{H}\right)\right)\end{aligned}\qquad(5\text{-}36)$$

其中 D 为特征向量 $X_{u,t}^{(s)}$ 的维数。将（5-36）式代入（5-29）式中，可以得到：

$$\begin{aligned}\frac{\partial g_i}{\partial c_h}&=\frac{1}{T_u^{(s)}}\sum_{t=1}^{T_u^{(s)}}\frac{\partial\ln b_i}{\partial c_h}=\frac{1}{T_u^{(s)}}\sum_{t=1}^{T_u^{(s)}}\frac{1}{b_i}\frac{\partial b_i}{\partial c_h}\\&=\frac{1}{T_u^{(s)}b_i}\sum_{t=1}^{T_u^{(s)}}\sum_{l=1}^{L}\frac{\partial w_{i,l}N_{i,l}(\mu_{i,l,d}\,|_{d=1}^{D},\{(\delta_{i,l,d})^2\}|_{d=1}^{D};X_{u,t}^{(s)})}{\partial c_h}.\end{aligned}\qquad(5\text{-}37)$$

由于篇幅关系，这里省略了更详细的展开。由于计算复杂度较高，在传统 DFE 研究中（Miyajima *et al*., 2001; Biem *et al*., 2001），通常出于简化算法的目的，假设 $\mu_{i,l,d}$ 和 $\delta_{i,l,d}$ 与 c_h 无关，即 $\frac{\partial\mu_{i,l,d}}{\partial c_h}=\frac{\partial\delta_{i,l,d}}{\partial c_h}=0$。

然而，实验表明，当利用特征参数向量更新完所有特征，再用新特征迭代估计 GMM 模型的参数时，性能提升有限。自动语音识别领域也存在类似现象（Povey *et al*., 2005），这一问题制约了 DFE 方法的发展，Biem 等人（2001）使用 k 最近邻分类器 kNN（k-Nearest Neighbour）来简化取代常用的 GMM 分类器。

为了解决这一问题，去掉上述假设，从 $\mu_{i,l,d}$ 和 $\delta_{i,l,d}$ 这样的 GMM 模型的参数中“间接”引入 c_h 的影响，即利用“间接”导数将模型参数迭代变化对 c_h 的影响考虑在内。具体来说，

$$\frac{\partial N_{i,l}(\mu_{i,l},\{(\delta_{i,l,d})^2\}|_{d=1}^{D};X_{u,t}^{(s)})}{\partial c_h}$$

$$=\frac{\exp\left\{\dfrac{\left(X_{u,t}^{(s)}-\mu_{i,l}\right)^T\left(X_{u,t}^{(s)}-\mu_{i,l}\right)}{2\prod_{d=1}^{D}(\delta_{i,l,d})^2}\right\}}{\sqrt[D]{2\pi}}\left[\frac{\partial\dfrac{1}{\prod_{d=1}^{D}\delta_{i,l,d}}}{\partial c_h}-\frac{1}{\prod_{d=1}^{D}\delta_{i,l,d}}\frac{\partial\dfrac{\left(X_{u,t}^{(s)}-\mu_{i,l}\right)^T\left(X_{u,t}^{(s)}-\mu_{i,l}\right)}{2\prod_{d=1}^{D}(\delta_{i,l,d})^2}}{\partial c_h}\right]. \tag{5-38}$$

显然 $\frac{\partial N_{i,l}(\mu_{i,l},\{(\delta_{i,l,d})^2\}|_{d=1}^{D};X_{u,t}^{(s)})}{\partial c_h}$ 式是关于 $\frac{\partial x_{u,t,d}^{(s)}}{\partial c_h}$、$\frac{\partial \mu_{i,l,d}}{\partial c_h}$、$\frac{\partial \delta_{i,l,d}}{\partial c_h}$ 的函数。

同时，

$$\frac{\partial \mu_{i,l,d}}{\partial c_h}=\frac{\partial}{\partial c_h}\frac{\sum_{u=1}^{U}\sum_{t=1}^{T_u}x_{u,t,d}^{(s)}}{\sum_{u=1}^{U}T_u}. \tag{5-39}$$

$$\frac{\partial \delta_{i,l,d}}{\partial c_h}=\frac{\partial}{\partial c_h}\sqrt{\frac{1}{\sum_{u=1}^{U}T_u}\cdot\sum_{u=1}^{U}\sum_{t=1}^{T_u}(x_{u,t,d}^{(s)}-\mu_{i,l,d})^2}. \tag{5-40}$$

可见（5-37）式中 $\frac{\partial g_i}{\partial c_h}$ 的是关于 $\frac{\partial x_{u,t,d}^{(s)}}{\partial c_h}$ 的函数。又有，

$$\frac{\partial x_{u,t,d}^{(s)}}{\partial c_h}=y_{u,t,h}^{(s)}\cos\left(\frac{d\pi(h+0.5)}{H}\right). \tag{5-41}$$

将（5-41）式逐步带回，就可以得到改进后的考虑了模型均值和方差随特征变化而得到的 $\frac{\partial g_i}{\partial c_h}$。由于篇幅原因在此省略了公式（5-37）的完整展开推导过程。

这样，将公式（5-37）的完整展开式带回公式式（5-22）～（5-24）就可以计算出 $L^{(s)}(C;\Lambda)$ 的偏导数，将其带回式（5-19）～（5-21）可以最终得到目标函数的梯度。

以上基于（5-36）式得到的仅是静态特征所对应的梯度。对于常用的差分特征，以一阶差分为例，即有，

$$z_{u,t,d}^{(s)} = \frac{\sum_{\theta=1}^{2}\theta(x_{u,(t+\theta),d}^{(s)} - x_{u,(t-\theta),d}^{(s)})}{2\sum_{\theta=1}^{2}\theta^2}. \tag{5-42}$$

$$\frac{\partial z_{u,t,d}^{(s)}}{\partial c_h} = \frac{1}{2\sum_{\theta=1}^{2}\theta^2}\left[\sum_{\theta=1}^{2}\theta(\frac{\partial x_{u,(t+\theta),d}^{(s)}}{\partial c_h} - \frac{\partial x_{u,(t-\theta),d}^{(s)}}{\partial c_h})\right]. \tag{5-43}$$

综上，就得到了实际应用的 GMM-UBM 声纹识别系统中，基于 MCE*MSV 准则进行 DFE 求时变意义下最优特征参数向量，即最优频带权值的计算方法。

5.4.4 准则的适用性

这种准则既可以用于对模型参数或特征参数单独进行训练，又可以用于对两种参数进行联合训练。

应用上述区分性训练的训练算法时，会使已有说话人模型之间的区分度提高。但当新增一个说话人模型的时候，很可能破坏原有模型的区分度。于是算法的应用场景需要特别注意。对于所有说话人通常都已知的闭集说话人辨认任务，应用模型参数的区分性训练后模型的区分度可以保持，故模型和特征参数的区分性训练都适于这类问题；对于不断新增说话人的确认任务，如果仅进行模型参数的区分性训练，在新增说话人模型的时候，模型组的区分度无法保持，所以这种应用不适合单独对模型参数进行训练，而更适合对特征参数进行优化。而进一步的，以上分析中所有可以应用模型参数区分性训练的场合都适合应用特征参数的区分性训练，故适于应用模型参数区分性训练的情况也都适于对模型和特征参数做联合优化以达到更好的系统性能。于是，在说话人识别的应用中，可优先对特征参数优化，然后根据不同应用条件选择性的将特征参数与模型参数进行联合训练。

5.5 基于GMM-UBM结构的快速梯度计算方法和弹性传播算法

在各种区分性训练中，获得参数梯度所需的大量计算一直是应用区分性训练主要的困难之一（Woodland and Povey, 2002）。5.2 节中介绍了使用基于 GPD 算法的随机优化方法来加速参数序列收敛，从而可以少计算参数梯度以实现区分性训练的加速。但是，需要注意的是，在 5.2.2 节 GPD 算法的应用条件中，条件（3）要求优化的目标函数具有唯一的极值点（最小值点）。这个条件对于优化目标为l_i线性函数的 MCE 准则来说，由于l_i在d_i上是单调递增的，故容易知道 MCE 准则的目标函数满足条件（3）。但是对于优化目标函数更为复杂的 MCE*MSV

准则来说，并不能保证唯一的极值点，因此不满足条件（3），不能直接应用 GPD 算法进行加速。为了解决这一问题，本节提出利用说话人识别技术中常用的 GMM-UBM 结构的特点对区分性训练中参数梯度进行快速计算的快速算法。

5.5.1 支配性高斯混合

在基于 GMM-UBM 的说话人识别系统中，UBM 模型为使用大量说话人数据训练而成的 GMM 模型，以体现说话人的共性信息；而说话人模型是在 UBM 模型的基础上，使用特定说话人的语音通过自适应技术得到的更突出该说话人个性信息的 GMM 模型，二者同构。在 GMM-UBM 系统中将语音在特定说话人模型上的对数似然度与它在 UBM 模型上的对数似然度做差就可以得到语音属于该说话人的置信度分数。在通过自适应技术得到说话人模型时，通常假设说话人之间的差异主要体现在 GMM 模型中各高斯混合的均值上，故常用 MAP 自适应技术对 UBM 模型中各个高斯混合的均值做说话人自适应。即说话人的个性信息体现在说话人模型中的每个高斯混合与其在 UBM 模型中对应高斯混合间的位置差异，即：

$$\begin{aligned} &Confidence(X;b_i) \\ &=\sum_{l=1}^{L} w_{i,l}N(\mu_{i,l},\Sigma_{i,l};X)-\sum_{l=1}^{L} w_{UBM,l}N(\mu_{UBM,l},\Sigma_{UBM,l};X). \end{aligned} \tag{5-44}$$

在 GMM 模型的所有高斯混合中，存在所谓支配性高斯混合（Reynolds, 2000），即某一语音帧在某几个高斯混合上的得分远大于（通常在 1000 倍以上）在模型中其它任何高斯成分上的得分。由于语音帧在这几个高斯混合上的得分在总得分中占据支配地位，称这几个高斯混合是该语音帧在该 GMM 模型中的支配性高斯混合。利用支配性高斯混合的原理，可以对声学模型的参数进行优化，或对在声学模型上的打分过程进行加速。如 Leggetter 和 Saraclar 等人在研究基于状态级发音模型算法时，当需要在某 GMM 中引入另一 GMM 的高斯混合进行模型扩充时，只在该 GMM 中增加另一 GMM 的支配性高斯混合，从而在保证扩充后模型性能的前提下尽可能多的降低新模型的复杂度（Leggetter and Woodland, 1995; Saraclar *et al*., 2000）；Gales 等人（1999）只计算语音帧在支配高斯混合上的得分以降低语音识别系统在搜索解码中的计算复杂度；Reynolds 等人（2000）利用 UBM 模型寻找输入语音帧的支配性高斯混合，只对相应的支配性高斯混合进行打分来加速，这也是当前 GMM-UBM 系统的标准做法。

5.5.2 快速梯度计算

利用支配性高斯混合的原理，通过近似来加速说话人识别系统中 μ、δ、w、C 等参数的计算。这种近似可以分为两个层面：

（1）公式（5-25）中，根据（5-27）式对某输入语音帧的 g_t 进行计算时，先计算 UBM 中该帧的支配性高斯混合，在每个说话人模型上进行计算时都只计算语音帧在支配性混合上的得分之和，以此近似语音帧在说话人模型上的总得分。

（2）公式（5-30）～（5-32）以及（5-37）中对各参数梯度进行计算时，只在模型的支配性高斯混合上进行并加和，作为此语音帧在模型上的准确梯度。但每个参数的梯度计算公式中除了包含与似然度计算相同的部分外，也有新增的部分。如：

$$\left[\frac{X_{u,t,d}^{(s)}-\mu_{i,l,d}}{\delta_{i,l,d}}\right]。$$

由于新增部分的存在，并不总能保证似然度大的高斯混合，其对应的梯度计算一定大。但考虑到支配性高斯混合上的得分通常比其它高斯混合上的得分大 1000 倍左右这样的事实，可以合理假设对大多数语音帧都有如下性质：语音帧在支配性高斯成分上的梯度也远大于它在非支配性高斯成分上计算得到的梯度，那么只需对支配性高斯混合而非 GMM 中的所有高斯混合进行梯度计算，从而加快计算速度。

本节使用的支配性高斯混合选择算法可以形式化表述如下。假设第 i 个说话人的 GMM 模型中的所有 L 个高斯混合已经按照各自似然度 $w_{i,l}N(\mu_{i,l},\Sigma_{i,l};X)$，$l=1,2,\ldots,L$ 递减的顺序进行了排序，则：

$$Dominant(i;P)=\{1,2,\ldots P\}. \qquad (5\text{-}45)$$

$Dominant(i;P)$ 为由说话人 i 的 P 个支配性高斯混合组成的支配性高斯混合的集合。通常 P 根据经验预先指定。显然，P 的选择是精度和速度之间的一种权衡；P 越大时快速算法中梯度近似的精度越高，但速度越慢。

由于 GPD 算法无法使用，那么将用标准的梯度计算公式对最优参数进行搜索，如下：

$$\Theta_{t+1}=\Theta_t-\varepsilon\nabla LV(C;\Lambda)|_{\Theta=\Theta_t}. \qquad (5\text{-}46)$$

5.5.3 弹性传播算法

式（5-46）中算法在沿梯度反方向搜索时每步搜索的步长难以确定。当

ε较小时，参数的收敛速度很慢，甚至可能陷入某个并不好的局部最优解；当ε较大时，搜索中可能过快的脱离包含较好局部最优解的区域，导致搜索不到局部最优解，如图 5.2 所示。

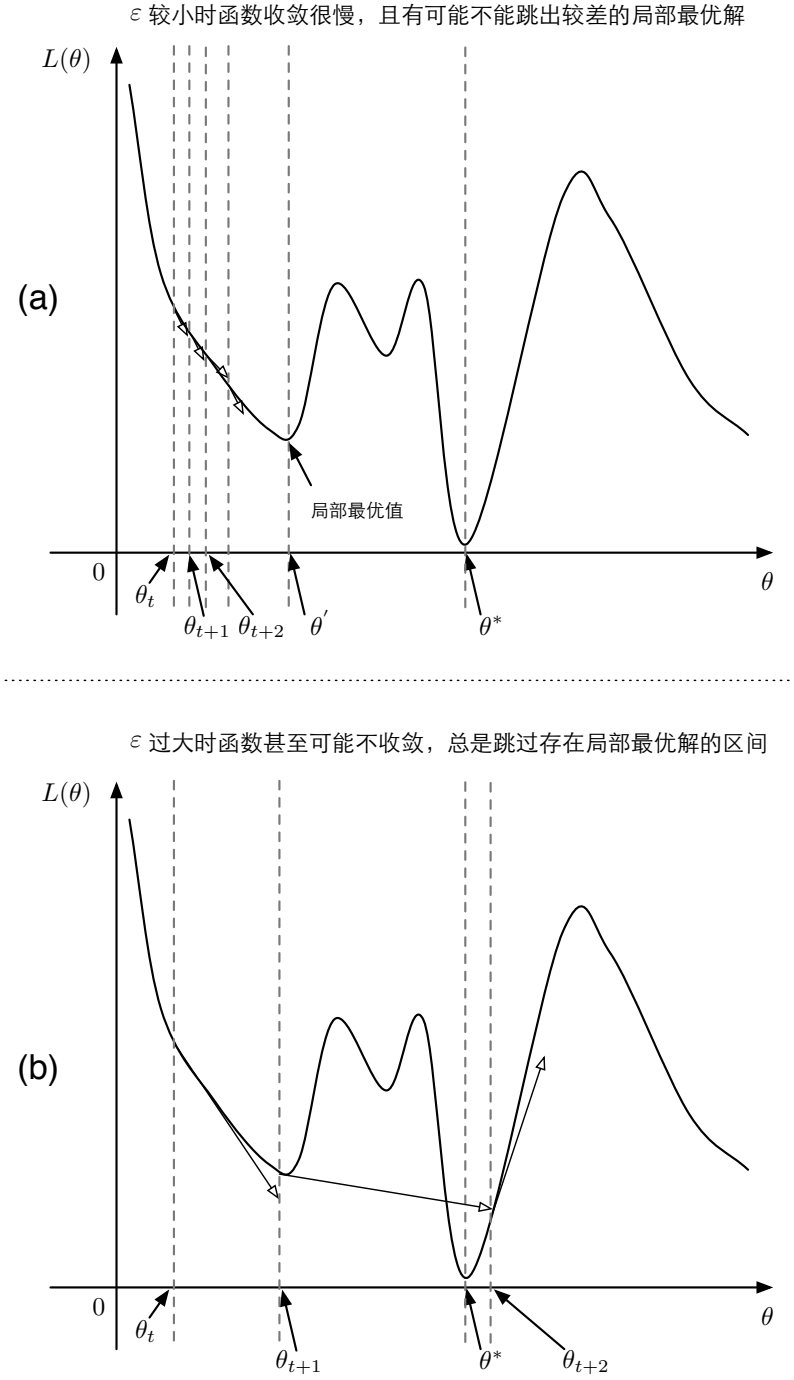

图5.2 搜索步长的影响示意图

另外，使用 GPD 或标准梯度下降算法时，虽然每个待优化参数的函数形式可能各不相同，但通常对所有参数选取统一的ε。这样可能导致ε对某些参数来说

过小，而对另一些参数来说却过大。一种较好的解决方法是对每个参数分别指定一个ε，并且让它们能够根据自身当前的搜索状况调整大小，这种梯度下降算法称为弹性传播 RProp（Resilient Propagation）算法（Riedmiller and Heinrich, 1993）。记参数θ第t次迭代时对应的参数为$\varepsilon_{\theta,t}$，则 RProp 算法可以写为：

$$\varepsilon_{\theta,(t+1)}=\begin{cases} s^{+}\varepsilon_{\theta,t}, \ if \ \dfrac{\partial LV(C_{t-1};\Lambda_{t-1})}{\partial\theta}\dfrac{\partial LV(C_t;\Lambda_t)}{\partial\theta}>0 \\ s^{-}\varepsilon_{\theta,t}, \ if \ \dfrac{\partial LV(C_{t-1};\Lambda_{t-1})}{\partial\theta}\dfrac{\partial LV(C_t;\Lambda_t)}{\partial\theta}<0 \end{cases}. \qquad (5\text{-}47)$$

$$\theta_{t+1}=\theta_t-\varepsilon_{\theta,t}\frac{\partial LV(C_t;\Lambda_t)}{\partial\theta}. \qquad (5\text{-}48)$$

即利用函数递增时一阶导数大于 0，函数递减时一阶导数小于 0 这一限制，依据参数在两次迭代间的导数的变化来控制步长。通常取$1<s^{+}$，$0<s^{-}<1$，即两次迭代间若导数符号相异，说明参数的目标函数刚刚发生了一次增减性变化，在两次迭代的参数值之间可能存在着局部最优解，于是应当缩小ε；而若导数符号一致，则说明参数正在目标函数的递增或递减区间内搜索，为了尽快使得目标函数到达局部最优解，应当增大ε。如图 5.3 所示。

需要注意的是，由于 GPD 算法对ε序列的收敛性有限制，故 RProp 算法只适合 MCE*MSV 准则使用的标准梯度下降算法。

5.6 实验

5.6.1 实验设置

实验数据的选取和划分、三角滤波器的个数、基线系统以及倒谱系数的配置与第 3 章的实验设置完全一致，详见 3.4.1 节，这里不再赘述。不同之处在于开发集数据（即第 3 章中用于训练 F-ratio 准则下的频带整体区分度的数据）用来进行区分性训练从而确定滤波器输出的权重（即前文所述对应频带的整体区分度）。

由于本章关注点为滤波之后的输出加权，因此本章实验中，滤波之前的频率弯折部分采用的是线性刻度变换。同第 3 章中 Weighting_F_ratio 的设置。

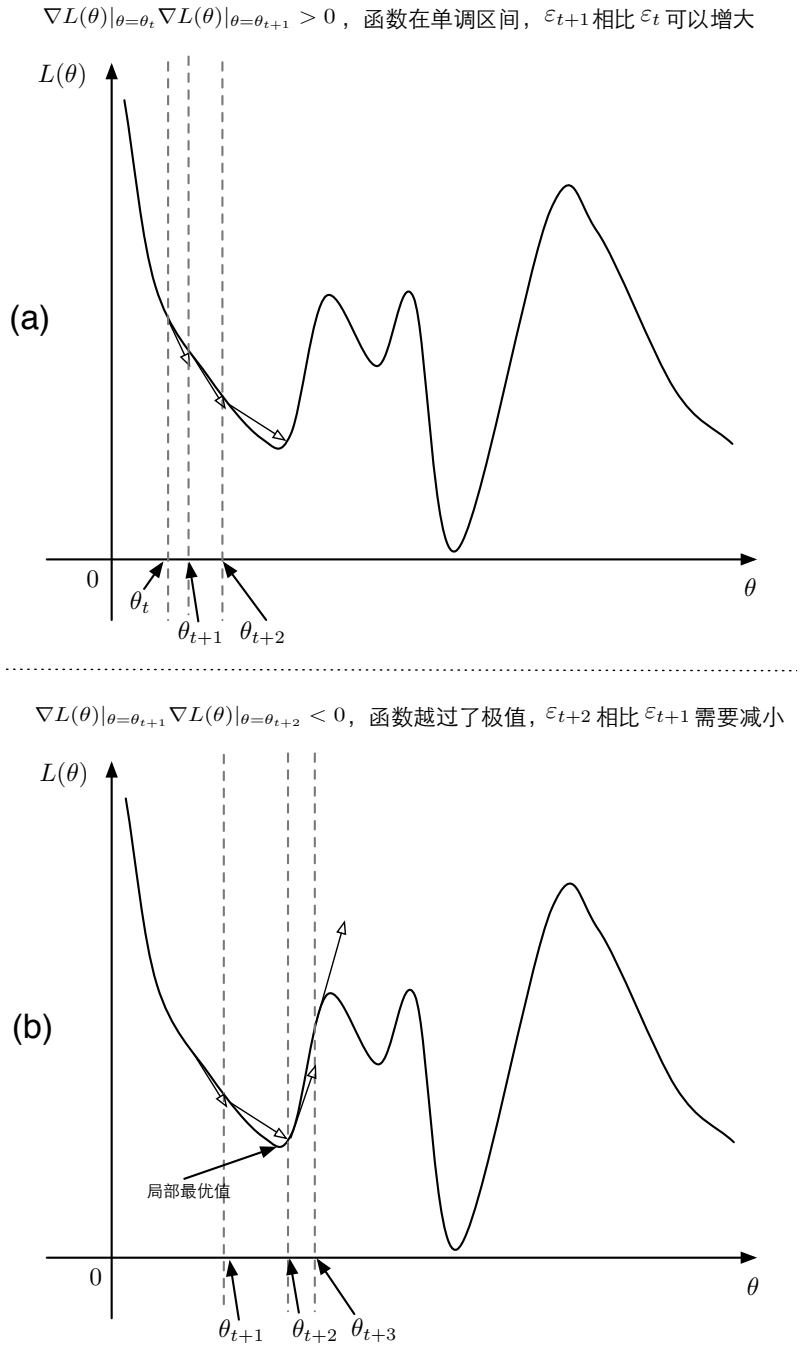

图5.3 RProp算法搜索步长弹性变化示意图

5.6.2 整体区分度

本章中模型的区分性训练算法主要基于说话人辨认系统的结果反馈进行，即MCE*MSV 准则中的错误率指的是首选的错误率。另外，考虑到本文的目的在于寻找既能强调说话人个性信息、又能弱化时间相关信息的特征参数，因此，本章在开发集数据上主要对使用区分性特征提取算法得到的特征参数（具体指滤波器

输出权重）进行实验，即利用 5.4.3 节中相关参数训练方法。

参数具体设置如下：取输出权重的初值采用了第 3 章中 F-ratio 准则下 Weighting_F_ratio 实验所采用过的输出权重。公式（5-1）中的η为 1.05；公式（5-2）中的α为 1.01；而 MCE*MSV 准则中的平衡因子，即公式（5-18）中的β，取 0.5，即同第 4 章均值*标准差的形式。

MCE*MSV 准则下训练得到的滤波器输出权重（即前文所述对应频带的整体区分度）如图 5.4 中折线所示。

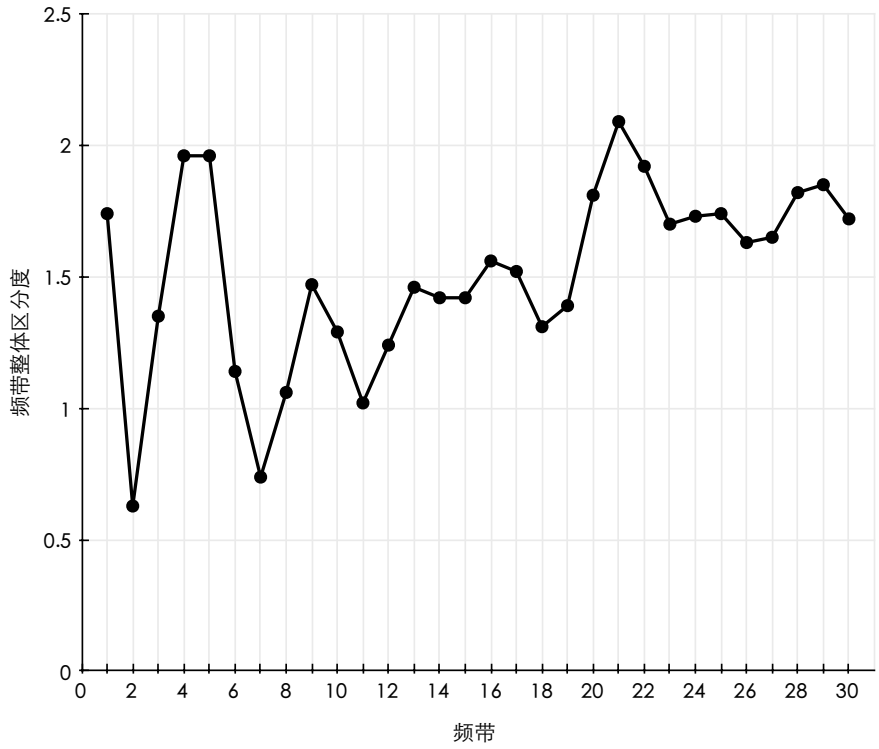

图5.4 MCE*MSV准则下的滤波器输出权重折线图

整体来看，与基于频带能量的 F-ratio 准则一致（见图 3.8），高频区域的区分度总体依然高于低频区域。抛开某些振荡点来看，折线整体呈现一种缓慢且稳定上升的趋势，与基于频带能量的 F-ratio 准则下呈现的快速攀升非常不同。MCE*MSV 准则下，高低频区分度的这种差距远没有基于频带能量的 F-ratio 准则下的数值差异那么大。巧合的是，虽然两种性能驱动准则的应用背景模式不同、本身数值间不具有可比性，但 MCE*MSV 准则下滤波器输出权重的这种波动模式类似于第 4 章中频率弯折下的性能驱动准则得到的整体区分度曲线。

5.6.3 加速性能

利用 5.5 节提到的支配性高斯混合思想以及 RProp 弹性传播算法，可以对 5.4.3 节中特征参数训练时的梯度计算进行加速。其中公式（5-47）中的 s^+ 取 1.2，而 s^- 取 0.5。

从加速比和区分性训练的性能两个方面，将原始的梯度计算方法和快速的梯度计算方法进行了对比，如表 5.1 所示。本实验选取的数据来自开发集，从每位说话人的每次录音会话中选取了五句。需要说明的是，这里的加速比是以特征参数作区分性训练时的实际收敛速度来计算的，而区分性训练的性能指的是在这批数据上得到的系统总体首选正确率。在所有的 1024 个高斯混合中选取似然度最大的 4 个作为支配性高斯混合，用它们的似然度的和近似 GMM 模型的似然度之和。

表5.1 原始梯度计算方法和快速梯度计算方法对比

梯度计算方法	性能	
	加速比	总体首选正确率(%)
原始梯度计算方法	1.00	74.2
快速梯度计算方法	2.99	73.9

从表 5.1 可以看出，在保证数据集上的首选正确率没有明显变化的前提下，快速的梯度计算较原始方法取得到了 2.99 倍的加速比。这说明之前关于支配高斯混合梯度也具有支配性这一假设合理，这对于参数调整目标函数复杂、收敛速度较慢的 MCE*MSV 准则具有特别重要的意义。

5.5.4 实验结果

基于图 5.4 中的权重折线对三角滤波器的输出做加权，提取倒谱特征（16 维倒谱加 16 维一阶差分）进行了说话人识别的实验，记为 Weighting_Pfm_drvn，并与基线系统的 MFCC 特征作了比较。如 5.5.2 节所述，MCE*MSV 准则是基于说话人辨认系统的性能，因此这两种倒谱系数特征在时变说话人识别系统中的性能对比亦是基于说话人辨认任务中的首选正确率。具体地，系统性能指标采用首选错误率的均值和标准差之积，见表 5.2。

表5.2 两种倒谱系数特征方法的辨认性能比较（%）

倒谱特征方法	首选错误率

	均值	标准差
MFCC	27.46	7.43
Weighting_Pfm_drvn	23.28	6.17

此外，为了基于频带能量和 F-ratio 准则下的 Weighting_F_ratio 方法作一对比，同样进行了说话人确认的实验。这三种倒谱系数特征在时变说话人识别系统中的确认性能对比见表 5.3。

表5.3 三种倒谱系数特征方法的确认性能比较（%）

倒谱特征方法	等错误率		下降率
	均值	标准差	均值*标准差
MFCC	8.80	1.86	--
Weighting_F_ratio	8.32	1.67	5.45
Weighting_Pfm_drvn	7.39	1.46	34.08

从 5.2 和 5.3 两张表中可以看出，无论是辨认性能还是确认性能，Weighting_Pfm_drvn 比基线系统的 MFCC 表现要优异不少。而对比滤波器输出加权模式下的两种准则：基于频带能量和 F-ratio 的准则与基于区分性训练的 MCE*MSV 准则，后者的表现超出许多，从某种意义上印证了与频率弯折相比，滤波器的输出加权需要一个更为复杂的参数确定过程。当然其中不能忽略的是，基于区分性训练的 MCE*MSV 准则需要大量的参数调整训练过程，其时间复杂度与基于频带能量和 F-ratio 的准则不是同一量级。

5.6 小结

本章主要探讨了滤波器输出加权方式下的性能驱动准则的应用。首先提出了利用区分性训练的思想进行时变说话人识别特征参数优化的思路，接着针对时变特点探讨了参数优化的具体准则及训练算法，同时阐述了特征与模型参数进行联合优化的方案，之后依据说话人识别模型的结构特点进行了加速处理。由于本文的目标在寻找更为鲁棒的特征，因此在本章的实验部分，主要针对特征参数进行了优化，取得了较好的效果。

第6章 总结和展望

6.1 论文工作总结

身份的数字化是当前信息化时代的一大特点，各式各样的密码也随时挑战着现代人的记忆极限，客观上推动了生物认证技术从实验室走向实际应用。国际生物集团 IBG（International Biometric Group）的最新报告表明，2014 年全球生物认证技术市场规模有望达到 90 亿美元（IBG, 2013）。而各项生物认证技术之中，据 Unisys 调查统计，在消费者偏好方面，说话人识别技术被排在首位；原因在于该技术有着天然优势，它对于语音的采集并不涉及到敏感的隐私信息，是一种非接触技术，易于依赖已有的电话网络等资源进行远程操作。因此说话人识别技术有着越来越广阔的应用前景。在其各方面的典型应用之中，声纹预留和声纹验证之间往往相隔一段时间，因而个体声纹随时间的变化就成为了说话人识别技术走向实用所无法回避的一个问题。本文对于说话人识别中的时变现象进行了初步的探索和研究，并提出了一系列针对时变现象的特征领域的鲁棒性算法，从实验的层面验证了算法的有效性，并从数据资源和特征提取两个方面为深入研究奠定了基础。

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）**建立了一个适合时变课题研究的声纹数据库**。数据是研究的根本和前提，而缺少一个合适的时变声纹库也是制约说话人识别领域时变研究的重要原因。在综合分析了现有的时变声纹资源之后，提出了时变声纹库设计的总体原则：尽最大可能保证时间是唯一改变因素。重点探讨了两个具体的设计原则：录音文本和时间间隔。为了尽量减少语音内容差异带来的影响，采用了固定的新闻语料作为录音文本，并要求说话人以朗读的方式完成，以尽量减少说话方式的变化。采用了梯度的时间间隔，最初语音采集比较频繁，之后间隔越来越长。此时变声纹库将通过 CCC 平台发布，以供研究人员使用。

（2）**提出了说话人确认系统时变鲁棒性的综合评价准则。**对于时变说话人确认系统而言，每次录音会话均存在一个等错误率；等错误率的均值代表了系统的平均等错误率水平，而等错误率的标准差则代表了系统性能随时间的变化性，因此本文以这一系列会话的等错误率的均值和标准差作为衡量系统的时变鲁棒性的重要评价指标，并定义两者的乘积为时变鲁棒性的综合评价准则。

（3）**提出了以 F-ratio 为中间准则计算频带区分度的时变鲁棒特征提取算法**。频带区分度是本文一以贯之的一个概念。从时变声纹库上的数据分析可见频带信息分布的时变现象，因此我们提出了频带整体区分度的概念，认为不同的频带对于时变说话人识别这一任务具有不同的区分度，而这一区分度需要综合考虑频带对于说话人个性信息的区分度，以及对于时间相关信息的区分度。于是，采用了基于频带能量和 F-ratio 的准则来计算各个频带的整体区分度，同时保证了与说话人个性信息的区分度正相关，而与时间相关信息的区分度负相关。确定了频带的区分度后，就可以设计时变鲁棒性算法，在特征提取的过程中给予不同频带不同的强调程度。主要探讨了两种方法：滤波之前的频率弯折以及滤波之后的输出加权。前者是通过加密或稀疏滤波器的设置来起到强调或弱化的效果，而后者则是通过直接加强或削弱对应频带在倒谱计算过程中的比重来实现。这两种算法分别记为 Warping_F_ratio 和 Weighting_F_ratio。

（4）**提出了基于性能驱动的频率弯折方法的特征提取算法**。在实际的系统中频带能量的区分度还会受到模型设计或参数优化等因素的影响，因而一种更直接的方式是从性能驱动的角度出发，利用实际的系统性能来评估频带的整体区分度。文中详细探讨了时变说话人识别任务的性能评价指标，提出了利用等错误率的均值和标准差之积来评价整体性能。在此基础上设计了针对频率弯折方式的性能驱动准则，即对于某一指定频带，保持其他所有频带分辨率不变，单独加强该频带，将系统的整体性能作为该频带的整体区分度指标。由此得到整体区分度曲线，并进行频率弯折，从而得到更为鲁棒的系统。这种算法记为 Warping_Pfm_Drvn。

（5）**提出了基于区分性训练的滤波器输出加权方法的特征提取算法**。（4）中的准则相当于利用识别结果进行了一次性能反馈，而本算法则是利用区分性特征提取的思想，给定滤波器输出权重一个初始序列，经过建模和打分过程，依据系统反馈的性能、通过一定的准则来调整输出权重，如此迭代若干次，直到找到一个性能比较好的权重序列。对于本算法而言，准则的选取尤为重要。针对时变说话人识别的特点，提出了 MCE*MSV 的准则，即迭代中优化的目标为使各次录音会话间错误率及其方差的函数最小化。最后得到最优输出权重即为频带的整体区分度。这种算法记为 Weighting_Pfm_Drvn。

以上提到的四种算法及基线系统（MFCC）的比较见表 6.1。

表6.1 各算法比较（%）

倒谱特征方法	等错误率		下降率
	均值	标准差	均值*标准差
MFCC	8.80	1.86	--
Warping_F_ratio	7.77	1.54	26.90
Weighting_F_ratio	8.32	1.67	5.45
Warping_Pfm_drvn	7.32	1.51	32.47
Weighting_Pfm_drvn	7.39	1.46	34.08

可见，四种算法相比于基线系统的 MFCC 都有或多或少的提升。基于频带能量和 F-ratio 的准则下，频率弯折方式要远好于滤波器输出加权的方式；而两种方式分别使用了针对自身的性能驱动准则来寻找更优的频带区分度曲线后，这种差距就大大缩小了，二者基本相当。从准则而言，两种性能驱动准则都得到了相比基于频带能量和 F-ratio 的准则更好一些的结果，这也从侧面反映了特征模块与模型模块间复杂的相互影响。当然，不能忽略的是，基于频带能量和 F-ratio 的准则，简单且易扩展，时间复杂度不高，与频率弯折方法相结合时，以一种十分经济的方式即取得了整体性能 26.9%的提升。相反，两种性能驱动准则均包含了大量的建模、打分过程，时间复杂度相对较高，这也限制了其扩展性。

6.2 下一步研究展望

本文对说话人识别中的时变现象进行了初步的研究，提出了一系列时变鲁棒性算法，在本文的时变声纹数据库 Chronos 上取得了一定的效果，但同时也有一些不足之处。针对这些不足之处，对于说话人识别中时变鲁棒性问题的进一步研究展望如下：

（1）时变声纹数据库 Chronos 目前涵盖的年龄段非常单一。考虑到长期录制的可操作性，选取了在校大学生（20 岁左右）来参加这个长达三年多的数据库录制项目。然而实验语音学的研究表明，每个年龄段的声音变化都各有不同，因此，一个更完善的时变声纹库需要涵盖更丰富的年龄段。考虑到儿童期和老年期人的声音状态变化剧烈，Chronos 的下一步扩展可先从中青年期入手，选取若干个年龄段，例如 30 岁左右、40 岁左右等，进行长期的语音录制。当然在数据库扩展的过程中，依然要遵循本文中所提到的总原则：尽最大可能保证时间是其中唯一变化因素。

（2）几种算法的叠加性研究没有涉及。本文提出了四种不同的时变鲁棒特征提取算法，但没有涉及到算法间的叠加。例如，如何将频率弯折方法与滤波器输出加权方法进行叠加，以进一步强调或弱化相关的频带，这是一个很有挑战性的课题。在这种复杂的情况下，又将如何确定频带的整体区分度，值得我们更深入的研究。

（3）时变鲁棒性研究仅着眼于以频带为核心的倒谱特征层面。 特征是模式识别问题的核心，所以本文将时变鲁棒性算法定位于从特征层面探索声纹的时变规律，然而从性能驱动准则的有效性可以看出，对于说话人识别这样一个典型的模式识别问题，特征模块和模型模块有着紧密的联系，且互相影响，将二者简单地割裂开来可能不会得到很好的效果。因此，下一步关于时变鲁棒性算法的研究，可从特征和模型相结合的层面进行，本文第 5 章中曾初步涉及过这种方式，可做更进一步的探讨，以期得到更好的效果。此外，本文以最有代表性的 MFCC 特征作为基线特征，并以其提取过程为例进行研究，而对于 LPCC、PLP、PLAR 等特征并没有过多涉及。

（4）时变鲁棒性研究依然建立在经典的 GMM-UBM 框架之下。如（2）中所述，由于定位在特征层面，暂时仅选用了经典的 GMM-UBM 说话人识别系统。然而近些年来，以 JFA 和 i-vector 为代表的说话人识别框架得到了越来越多的重视，并在很多领域，如跨信道问题、短语音问题等，取得了很好的效果。这也是今后时变课题的一个研究方向。

（5）时变鲁棒性算法仅在本文的时变声纹库 Chronos 上进行了实验。由于课题的特殊性和现有资源的限制，本文提出的鲁棒性算法仅在 Chronos 上进行了实验，取得了一定的效果。但在其他录制条件下的声纹库上是否依然存在文中所描述的时变规律，还需要在数据支持基础上的进一步验证。

参考文献

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

Beigi H, 2009. Effects of time lapse on speaker recognition results. Proceedings of 16th International Conference on Digital Signal Processing, 1-6.

Beigi H, 2010. Foundations of speaker recognition. New York: Springer.

Biem A and Katagiri S, 1993. Feature extraction based on minimum classification error/generalized probabilistic descent method. Proceedings of the 17th International Conference on Audio, Speech, and Signal Processing, ICASSP, Montreal, Canada, 275-278.

Biem A, Katagiri S, McDermott E, *et al*., 2001. An application of discriminative feature extraction to filter-bank-based speech recognition. IEEE Transactions on Speech and Audio Processing, 9(2):96-110.

Bimbot F, Bonastre J, Fredouille C, *et al*., 2004. A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing, 4:430-451.

Bishop C M, 2007. Pattern recognition and machine learning. New York: Springer.

Boersma P, 2002. Praat, a system for doing phonetics by computer. Glot International, 5(9/10):341-345.

Bonastre J, Bimbot F, Boe L, *et al*., 2003. Person authentication by voice: a need for caution. Proceedings of Eurospeech 2003, Geneva, 33-36.

Brandschain L, Graff D, Cieri C, *et al*., 2010. Greybeard - voice and aging. Proceedings of 7th Conference on International Language Resources and Evaluation, LREC'10.

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

Campbell J and Higgins A, 1994. YOHO speaker verification. Linguistic Data Consortium.

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

CCC, 2004. CCC Homepage [online]. Available from: http://www.d-ear.com/CCC/ [Accessed 10 April 2013].

Chou W, Juang B-H, and Lee C-H, 1992. Segmental GPD training of HMM based speech recognizer. Proceedings of the 16th International Conference on Audio, Speech, and Signal Processing, ICASSP, San Francisco, USA, 473-476.

Cieri C, Walt A, Campbell J P, *et al*., 2006. The mixer and transcript reading corpora: resources for multilingual, crosschannel speaker recognition research. Proceedings of 5th International Conference on Language Resources and Evaluation, LREC.

Cieri C, Corson L, Graff D, *et al*., 2007. Resources for New Research Directions in Speaker Recognition: the mixer 3, 4 and 5 corpora. Proceedings of Interspeech 2007, Antwerp.

CLDC, 2004. CLDC Homepage [online]. Available from: http://www.chineseldc.org [Accessed 10 April 2013].

Cole R, Noel M, and Noel V, 1998. The CSLU speaker recognition corpus. Proceedings of International Conference of Spoken Language Processing, ICSLP 1998, 3167-3170.

Cumani S, Brummer N, Burget L, *et al*., 2011. Fast discriminative speaker verification in the i-vector space. Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, 4852-4855.

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

Dehak N and Kenny P, 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. Proceedings of Interspeech 2009, Brighton, UK.

Dehak N, Kenny P, Dehak R, *et al*., 2011. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing, 19(4):788-798.

Dobrisek S, Mihelic F, and Pavesic N, 1999. Acoustic modeling of phone transitions: biphones and diphones - what are the differences. Proceedings of Eurospeech 1999, 1307-1310.

ELRA, 2008. ELRA Homepage [online]. Available from: http://www.elra.info [Accessed 10 April 2013].

Fu Q, Zhao Y and Juang B-H, 2012. Automatic speech recognition based on non-uniform error criteria. IEEE Transactions on Audio, Speech, and Language Processing, 20(3): 780-793.

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

Furui S, 1997. Recent advances in speaker recognition. Pattern Recognition Letters, 18(9):859-872.

Gales M J F, Knill K M and Young S J, 1999. State-based Gaussian selection in large vocabulary continuous speech recognition using HMMs. IEEE Transactions on Speech and Audio Processing, 7(2): 152-161.

Gauvin J L and Lee C-H, 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, 2(2):291-298.

Glembek O, Burget L, Matejka P, *et al*., 2011. Simplification and optimization of i-vector extraction. Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, 4516-4519.

Godfrey J J, Holliman E C, and McDaniel J, 1992. SWITCHBOARD: telephone speech corpus for research and development. Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, ICASSP 1992, 1:517-520.

Godin K W and Hansen J H L, 2010. Session variability contrasts in the MARP corpus. Proceedings of 11th Annual Conference of the International Speech Communication Association, 298-301.

Goh K I, Cusick M E, Valle D, *et al*., 2007. The human disease network. Proceedings of the

National Academy of Sciences, 104(21):8685-8690.

GSK, 2006. GSK Homepage [online]. Available from: http://www.gsk.or.jp/index_e.html [Accessed 10 April 2013].

Hartman D, 1979. The perceptual identity and characteristics of aging in normal male adult speakers. Journal of Communication Disorders, 12:53-61.

Hartman D and Danhauer J L, 1979. Perceptual features of speech for males in four perceived age decades. Journal of the Acoustical Society of America, 59:713-715.

Hebert M, 2008. Text-dependent speaker recognition. Springer Handbook of Speech Processing, Berlin: Springer-Verlag.

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

Horri Y and Ryan W, 1981. Fundamental frequency characteristics and perceived age of adult male speakers. Folia Phoniatrica, 33:227-233.

Huang X-D, Acero A, and Hon H, 2001. Spoken language processing: a guide to theory, algorithm and system development. New Jersey: Prentice Hall, 419-426.

IBG, 2013. BMIR 2009-2014 [online]. Available from: http://ibgweb.com/products/reports/bmir-2009-2014 [Accessed 10 April 2013].

Juang B-H, Levison S E, and Sondhi M M, 1986. Maximum likelihood estimation for multivariate mixture observations of Markov Chains. IEEE Transactions on Information Theory, 32(2): 307-309.

Juang B-H, Wu C, and Lee C-H, 1997. Minimum classification error rate methods for speech recognition. IEEE Transactions on Speech and Audio Processing, 5(3):257-265.

Kato T and Shimizu T, 2003. Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence preserving connected digit patterns. Proceedings of ICASSP 2003, Hong Kong, 57-60.

Kelly F, Drygajlo A, and Harte N, 2012. Speaker verification with long-term ageing data. Proceedings of 5th IAPR International Conference on Biometrics, New Delhi.

Kelly F and Harte N, 2011. Effects of long-term ageing on speaker verification. Biometrics and ID Management: Lecture Notes in Computer Science, Berlin/Heidelberg: Springer, 6583:113-124.

Kelly F, Dyrgajo A, and Harte N, 2013. Speaker verification in score-ageing-quality classification space. Journal of Computer Speech and Language.

Kenny P, 2005. Joint factor analysis of speaker and session variability: theory and algorithms. Technical report CRIM-06/08-13 Montreal, CRIM.

Kenny P, 2010. Bayesian speaker verification with heavy tailed priors. Proceedings of IEEE Odyssey Speaker and Language Recognition Workshop 2010.

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

Kenny P, Boulianne G, Ouellet P, *et al*., 2007. Speaker and session variability in GMM-based

speaker verification. IEEE Transactions on Audio, Speech and Language Processing, 15(4):1448-1460.

Kenny P, Ouellet P, Dehak N, *et al*., 2008. A study of inter-speaker variability in speaker verification. IEEE Transactions on Audio, Speech and Language Processing, 7:980-988.

Kersta L G, 1962. Voiceprint recognition. Nature, 4861:1253-1257.

Kharroubi J, Petrovska D D, and Chollet G, 2001. Combining GMMs with support vector machines for text independent speaker verification. Proceedings of the European Conference on Speech Communication and Technology, Eurospeech, Aalborg, Denmark, 1757-1760.

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

Kinnunen T and Li H, 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Communication, 52:12-40.

Kunzel H J, 1994. Current approaches to forensic speaker recognition. ESCA workshop on Automatic Speaker Recognition, Identification and Verification, 135-141.

Lamel L and Gauvin J, 2000. Speaker verification over the telephone. Speech Communication, 31:141-154.

Lawson A D, Stauffer A R, Cupples E J, *et al*., 2009a. The multi-session audio research project (MARP) corpus: goals, design and initial findings. Proceedings of Interspeech 2009, Brighton, UK, 1811-1814.

Lawson A D, Stauffer A R, Cupples E J, *et al*., 2009b. Long-term examination of intra-session and inter-session speaker variability. Proceedings of Interspeech 2009, Brighton, UK, 2899-2902.

LDC, 1992. LDC Homepage [online]. Available from: http://www.ldc.upenn.edu [Accessed 10 April 2013].

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

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

Leggetter C J, 1995. Improved acoustic modeling for HMMs using linear transformation. Phd thesis. Cambridge University.

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

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

Lei H and Lopez E, 2009. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. Proceedings of Interspeech 2009, Brighton, UK, 2323-2326.

Li J and Lee C-H, 2007. Soft margin feature extraction for automatic speech recognition. Proceedings of the 10th European Conference on Speech Communication and Technology (Interspeech-Eurospeech), Antwerp, Belgium, 30-33.

Li J, Zheng F, and Wu W-H, 2004. Context dependent initial/final acoustic modeling for Chinese continuous speech recognition. Journal of Tsinghua University (Sci & Tech), 24(1):61-64.

Li J, Zheng F, Xiong Z-Y, *et al*., 2003. Construction of large-scale Shanghai Putonghua speech corpus for Chinese speech recognition. Proceedings of Oriental-COCOSDA, 62-69.

Li J, Zheng F, Zhang J-Y, *et al*., 2001. The definition and extension of the question set for decision tree based state tying in Chinese speech recognition. Proceedings of International Conference on Chinese Computing 2001, 106-110.

Li M, Zhang X, Yan Y-H, *et al*., 2011. Speaker verification using sparse representations on total variability i-vectors. Proceedings of Interspeech 2011, 2729-2732.

Liang C-Y, Zhang X, Yang L, *et al*., 2012. Discriminative decision function based scoring method in joint factor analysis for speaker verification. Proceedings of Interspeech 2012.

Lin T and Wang L-J, 1991. A course book of phonetics. Beijing University Press.

Linville S E, 2004. The aging voice. The American Speech-Lanugage-Hearing Association (ASHA) Leader, 12-21.

Linville S E and Fisher H B, 1985. Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. Journal of the Acoustical Society of America, 78:40-48.

Lomax R G and Hahs-Vaughn D L, 2007. Statistical concepts: a second course. Mahwah: Lawrence Erlbaum Associates.

Lu X-G and Dang J-W, 2007. Physiological feature extraction for text independent speaker identification using non-uniform subband processing. Proceedings of ICASSP 2007, Hawaii, 461-464.

Lu X-G and Dang J-W, 2008. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Communication, 50:312-322.

Markel J and Davis S, 1979. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP, 27(1):74-82.

Matsui T and Furui S, 1992. Comparison of text-independent speaker recognition methods using VQ distortion and discrete/continuous HMMs. Proceedings of the International Conference of Acoustics, Speech, and Signal Processing, ICASSP 1992, 2:157-160.

McDermott E and Hazen T J, 2004. Minimum classification error training of landmark models for real-time continuous speech recognition. Proceedings of the 28th International Conference on Audio, Speech, and Signal Processing, ICASSP, Montreal, Canada, 937-940.

Miller D, Cieri C, and Walker K, 2001. Switchboard Cellular Resources for Speaker Recognition. NIST Speaker Recognition Workshop 2001, Maritime Institute of Technology and Graduate Studies, Linthicum MD.

Mitchell M, 1999. An introduction to genetic algorithm. Cambridge: MIT Press.

Miyajima C, Watanabe H, Tokuda K, *et al*., 2001. A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction. Speech Communication, 35:203-218.

Mueller P B, 1989. Voice characteristics of centenarian subjects. The 21st Congress of the International Association of Logopedics and Phoniatrics, Prague.

Nakagawa S, Zhang W, and Takahashi M, 2004. Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM. Proceedings of ICASSP 2004, Montreal, 81-84.

NII, 2007a. AWA-LTR [online]. Available from: http://research.nii.ac.jp/src/en/AWA-LTR.html [Accessed 10 April 2013].

NII, 2007b. NII-SRC [online]. Available from: http://research.nii.ac.jp/src/en/index.html [Accessed 10 April 2013].

NIST, 1995. NIST SRE [online]. Available from: http://www.itl.nist.gov/iad/mig/tests/sre/ [Accessed 10 April 2013].

NIST, 2010. NIST SRE [online]. Available from: http://www.itl.nist.gov/iad/mig/tests/sre/2010/ [Accessed 10 April 2013].

Orlikoff R F, 1990. The relationship of age and cardiovascular health to certain acoustic characteristics of male voices. Journal of Speech, Language and Hearing Research, 33(3):450-457.

Povey D, Kingsbury B, Mangu L, *et al*., 2005. FMPE: discriminatively trained features for speech recognition. Proceedings of the 29th International Conference on Audio, Speech, and Signal Processing, ICASSP, Orlando, USA, 105-108.

Povey D and Woodland P C, 2002. Minimum phone error and I-smoothing for improved discriminative training. Proceedings of the 26th International Conference on Audio, Speech, and Signal Processing, ICASSP, Orlando, USA, 105-108.

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

Ptacek P H and Sander E K, 1966. Age recognition from voice. Journal of Speech and Hearing Research, 9:273-277.

Rabiner L and Juang B-H, 1993. Fundamentals of speech recognition. Signal Processing Series. Englewood Cliffs, NJ: Prentice Hall.

Reubold U, Harrington J, and Kleber F, 2010. Vocal aging effects on F0 and the first formant: a longitudinal analysis in adult speakers. Speech Communication, 52:638-651.

Reynolds D A, Quatieri T F, and Dunn R B, 2000. Speaker verification using adapted Gaussian

mixture models. Digital Signal Processing, 10:19-41.

Rhodes R, 2011. Changes in the voice across the early adult lifespan. The International Association of Forensic Phonetics and Acoustics, IAFPA 2011.

Riedmiller M and Heinrich B, 1993. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. Proceedings of IEEE International Conference on Neural Networks, San Francisco, California, USA, 586-591.

Robbins H and Monro S, 1951. Approximation method. Annals of Mathematical Statistics, 22(3): 400-407.

Rodgers J L and Nicewander W A, 1988. Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1):59-66.

Rose P, 2002. Forensic speaker identification. London: Taylor & Francis.

Ryan W and Burk K, 1974. Perceptual and acoustic correlates of aging in the speech of males. Journal of Communication Disorders, 7:181-192.

Saraclar M, Nock H, and Khudanpur S, 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language, 13(4): 137-160.

Schluuter R, Macherey W, Muuller B, *et al*., 2001. Comparison of discriminative training criteria and optimization methods for speech recognition. Speech Communication, 34: 287-310.

Segre R, 1971. Senescence of the voice. Eye, Ear, Nose, and Throat Monthly, 50:223-227.

Senoussaoui M, Kenny P, Dehak N, *et al*., 2010. An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. Proceedings of IEEE Odyssey Speaker and Language Recognition Workshop 2010, 28-33.

Senoussaoui M, Kenny P, Brummer N, *et al*., 2011. Mixture of PLDA models in i-vector space for gender independent speaker recognition. Proceedings of International Conference on Speech Communication and Technology 2011.

Sha F, 2006. Large Margin Training of Acoustic Models for Speech Recognition. PhD Dissertation. University of Pennsylvania.

Shipp T and Hollien H, 1969. Perception of the aging male voice. Journal of Speech and Hearing Research, 12:703-710.

Sinha R, Tranter S, Gales M, *et al*., 2005. The Cambridge University March 2005 speaker diarization system. Proceedings of the European Conference on Speech Communication and Technology, 2437-2440.

Snyman J A, 2005. Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms. New York: Springer.

Soong F, Rosenberg A E, Rabiner L R, *et al*., 1985. A vector quantization approach to speaker recognition. Proceedings of ICASSP 1985, Florida, 10:387-390.

Sorenson H W and Alspach D L, 1971. Recursive Bayesian estimation using Gaussian sums. Automatica, 7: 465-479.

Stathopoulos E T, Huber J E, and Sussman J E, 2011. Changes in acoustic characteristics of the

voice across the life span: measures from individuals 4-93 years of age. Journal of Speech, Language, and Hearing Research, 54:1011-1021.

Stevens K N, 1998. Acoustic phonetics. Cambridge, MA: MIT Press.

Stigler S M, 1989. Francis Galton's account of the invention of correlation. Statistical Science, 4(2):73-79.

Tranter S, Yu K, Evermann G, *et al*., 2004. Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, 433-437.

Valtchev V, Odell J, Woodland P, *et al*., 1997. MMIE training of large vocabulary recognition systems. Speech Communication, 22:303-314.

Vapnik V N, 1998. Statistical learning theory. New York: Wiley-Interscience.

Vogt R and Sridharan S, 2006. Experiments in session variability modeling for speaker verification. Proceedings of International Conference on Acoustic, Signal, Speech and Processing, ICASSP 2006, 897-900.

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

Wolf J J, 1972. Efficient acoustic parameters for speaker recognition. Journal of Acoustic Society of America, 51(6):2044-2056.

Woodland P C and Povey D, 2002. Large scale discriminative training of hidden Markov models for speech recognition. Computer Speech and Language, 16:25-47.

Xiong Z-Y, Zheng F, Li J, *et al*., 2003. Au automatic prompting texts selecting algorithm for di-Ifs balanced speech corpus. Proceedings of National Conference on Man-Machine Speech Communications, NCMMSC 2003, 252-256.

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

Young S, Evermann G, Kershaw D, *et al*., 2002. The HTK book (for HTK version 3.2). Cambridge University Engineering Department.

Zhang C and Hansen J H L, 2007. Analysis and classification of speech mode: whispered through shouted. Proceedings of Interspeech 2007, Antwerp, 2289-2292.

Zhou X-H, Garcia-Romero D, Duraiswami R, *et al*., 2011. Linear versus Mel frequency cepstral coefficients for speaker recognition. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Hawaii, 559-564.

陈文翔，杨莹春，2010. 声纹漂移现象初探. 第九届中国语音学学术会议.

龚光鲁，钱敏平，2007. 应用随机过程教程. 北京：清华大学出版社.

侯丽珍，2010. 嗓音年龄变化特点及其解剖生理机制的探讨. 国外医学耳鼻咽喉科学分册, 5:287-291.

黄挺，2011. 情感说话人识别中的基频失配及其补偿方法研究[博士学位论文]. 杭州：浙江大

学.

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

陆伟，2008. 基于缺失特征的文本无关说话人识别鲁棒性研究[博士学位论文]. 合肥：中国科技大学.

单振宇，2010. 情感说话人识别及其解决方法的研究[博士学位论文]. 杭州：浙江大学.

单振宇，杨莹春，2005. 声纹打卡系统. 第八届全国人机语音通讯学术会议，NCMMSC，565-568.

甄斌，吴玺宏，刘志敏，迟惠生，2001. 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），37(3):371-378.

周顺忠，邢志忠，2006. 比约肯谈实验数据的重要性. 现代物理知识, 17(2):66-67.

致　谢

衷心感谢我的导师郑方研究员六年来的悉心指导和关怀。郑老师严谨治学、坦荡做事、平易近人，他的言传身教将使我受益终生。读博期间，在学术上经历了很多挫折，但郑老师总是能给我以细致的指导，且自始至终都给了我很多的鼓励和极大的包容。在此，谨向恩师致以最诚挚的谢意！

感谢我硕士期间的导师徐明星副教授。徐老师手把手带领当时对信号一无所知的我进入说话人识别研究的领域，无论科研还是生活中，徐老师永远是亲切且耐心的春风化雨。谢谢徐老师一直以来的关心。

感谢我在 CSLT 的第一位指导教师夏云庆副研究员。初到 CSLT，夏老师给了我很大的帮助，在夏老师指导下工作的半年也让我了解了什么是做研究。谢谢夏老师。

感谢邬晓钧老师以及 CSLT 所有关心帮助过我的老师，他们的学识和品格永远值得我学习。

感谢 VPR 组的张利鹏师兄、王刚师兄以及张陈昊、罗灿华 、龚成、别凡虎、王军等同学，他们给予了我很多学习与工作上的支持和帮助，一起讨论科研，一起探讨生活，谢谢他们给了我家的感觉。

感谢实验室所有同学，他们共同营造了温暖的氛围。

感谢参与时变声纹库数据采集的九字班 50 多位同学。谢谢他们坚持下来这个三年的录音项目，很抱歉无法在此一一写出每位同学的名字，谨向他们致谢！声纹库的录制还得到了唐国瑜、陈丽欧、李超及龚成的通力协助，不胜感激。

最后感谢我的家人，无私的爱和无条件的支持，才能让我轻松地走到现在。谢谢你们！

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

1985 年 8 月 19 日出生于山东省菏泽市定陶县。

2002 年 9 月考入北京语言大学计算机科学与技术系，2006 年 7 月本科毕业并获得工学学士学位。

2007 年 9 月考入清华大学计算机系攻读计算机应用博士至今。

在学期间发表的学术论文

[1] Linlin Wang, Xiaojun Wu, Thomas Fang Zheng, and Chenhao Zhang. An investigation into better frequency warping for time-varying speaker recognition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, Los Angeles, 2012. (EI 会议，检索号 20131016079194)

[2] Linlin Wang, Thomas Fang Zheng, Chenhao Zhang, and Gang Wang. Discrimination-emphasized mel-frequency-warping for time-varying speaker recognition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, p 731-734, Xi'an, 2011. (优秀学生论文)(EI 会议，检索号 20124015499682)

[3] Linlin Wang and Mingxing Xu. SDBM-based speaker recognition for speaking style variations. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, p 744-747, Xi'an, 2011. (EI 会议，检索号 20124015499685)

[4] Linlin Wang and Thomas Fang Zheng. Creation of time-varying voiceprint database. Oriental-COCOSDA (the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques), Kathmandu, 2010.

[5] Linlin Wang, Lipeng Zhang, and Mingxing Xu. Score normalization-based speaking-style variation robust speaker recognition. Qinghua Daxue Xuebao/Journal of Tsinghua University, v 49, n SUPPL 1, p 1278-1282, 2009. (EI 期刊，检索号 20103413174610)

[6] Yunqing Xia, Linlin Wang, Kam-Fai Wong, and Mingxing Xu. Sentiment vector space model for lyric-based song sentiment classification. 46th Annual Meeting of the Association for Computational Linguistics: Human

Language Technologies, ACL-08: HLT, p 133-136, Ohio, 2008. (EI 会议, 检索号: 20121714962092)

[7] Yunqing Xia, Linlin Wang, and Kam-Fai Wong. Sentiment vector space model for lyric-based song sentiment classification. International Journal of Computer Processing of Oriental Languages, v 21, n 4, p 331-345, 2008.