摘 要

随着信息技术的迅速发展，人机交互技术的不断普及，说话人识别（Speaker Recognition, SR）以其独特的方便性、经济性和安全性等优势受到了越来越多人的关注，在信息安全等领域的应用也逐渐增加。同时，随着嵌入式系统在处理速度、存储能力、功耗和体积等方面取得突破性的进展，嵌入式说话人识别系统逐渐成为语音识别技术面向实际应用的一个重要发展趋势。然而将与文本无关的说话人识别系统应用到嵌入式设备上，依旧面临着嵌入式设备计算速度、存储能力等资源受限问题和背景噪声、跨信道等算法鲁棒性问题，影响嵌入式说话人识别系统的精度和实际应用效果。

针对上述问题，本文在嵌入式说话人识别系统的运行效率和识别性能两方面进行了研究和改进。主要内容包括：

为了提高系统的运行效率使其能在嵌入式设备上得以应用，引入了一种快速算法——非线性分段（Non-Linear Partition, NLP）算法。由于该算法基于距离累积的分段规则对语音中的微小干扰鲁棒性较差，本文引入了改进的 NLP 算法，采用绝对值距离替代平方和距离，并引入马氏距离作为新的分段规则。实验证明，改进后的 NLP 算法使得分段的结果更加稳定。实验结果表明，采用改进的 NLP 算法相对基线系统整体性能提升 20.22%

为了提高系统的识别性能以及增强系统鲁棒性，针对不同的人发音习惯不同的现象，采用了一种基于基频曲线的特征来着重捕捉说话人较长时间的韵律信息。通过研究现有的一些融合方案，在基线系统的框架上进行改进，提出了一种在分数层上进行分类器融合的方法。该方法更加充分地利用训练用的语料，使得训练用的语料除了用来训练前端数学模型之外，还用来调整后端模型（支持向量机）的超参数，且获得了好的整体辨识结果。提出的多重特征融合的方法得到了最高的辨识率，相对基线系统整体性能提升了 47.57%。

关键词： 嵌入式平台，说话人识别，文本无关，支持向量机，分类器融合

Abstract

With the rapid development of Information Technology (IT) and the spread of human-computer interaction, the speaker recognition technology attracts more and more attention due to its convenience，economy and security, and its application increases dramatically in the field of information security. Meanwhile, embedded speaker recognition system becomes an important tendency in Application of the speech recognition technology for the breakthrough in the progress of embedded system in power consumption. However, text-independent speaker recognition system in embedded devices still facing many problems such as low computing speed, limited storage capacity, background noise and robustness in cross-channel, which can reduce accuracy of embedded speaker recognition system and its practical effect.

To solve these problems，this thesis researches and improves the efficiency and recognition rate in embedded speaker recognition systems. The main contents include:

This thesis introduces a fast algorithm--Non-Linear Partition, to improve the efficiency of the system improve the system operating efficiency. Because the poor robustness of the segmentation rule based on the distance accumulated on the small interference in the voice, it employ absolute value of the distance instead of the square and distance as well as Mahalanobis distance as a new segmentation rule to improve NLP. Experimental results show that the improved NLP algorithm makes the segmentation results are more stable and relative to baseline system to enhance the overall performance of the 20.22%.

According to the different pronunciation habits of different phenomena, this paper uses a type of speech feature to focus on capturing the speaker longer prosodic information based on the pitch contour in order to improve the performance of the system and enhance system robustness. We made a review of several existing fusion program and improve the baseline system framework so that this design can use multiple features, and the identification results of sub-classifiers are fused on score level. The training corpus are not only used to train the models, but also used to tune the parameters which will be further used in the Support Vector Machine on the back-end, in order to take full advantage of training corpus and get a good overall recognition result. Experimental results show that the multi-feature fusion system to get the highest recognition rate compare to the baseline system to enhance the

overall performance of the 47.57.

Key words: Embedded platform，Speech recognition,，Text-independent，Support vector machine，Classifier fusion

目 录

第1章 绪论

1.1 引言

说话人识别又称声纹识别，是利用人体生物特征进行身份认证的一种技术，是目前最为方便与直接的一种识别技术[1]。当前，对说话人识别技术的研究大多是基于 PC 机平台上、安静的实验室环境下进行的。然而，随着移动通信的蓬勃发展和便携式设备的推陈出新，说话人识别技术今后将更多应用于嵌入式设备上。目前嵌入式设备上实现的说话人识别系统大多是基于文本相关的简单命令词识别，实用性差和灵活性低等特点制约了其在实际应用中的推广。

本文通过对现有的面向嵌入式系统的文本无关的说话人识别技术进行研究，分析影响嵌入式说话人识别系统性能的主要原因。从提高系统运行效率的方面考虑而引入了 NLP 算法，从提高系统鲁棒性的方面考虑而采用了多特征融合的方法。并选取说话人识别技术中常用的等错误率（Equal Error Rate，EER）对系统性能进行评价，证明了本文方法的合理性和有效性。

1.2 课题背景和研究意义

语言是人与人交流最为自然和方便的工具。随着信息时代的步入，人类和机器的交互越来越频繁和迫切，要求机器也能听懂人类的语言，并能自然地同人类进行交流。说话人识别技术作为机器理解和识别人类语言的一个分支学科，旨在能从人类的语音信息里找到能表征说话人身份的信息，能对说话人自身的身份进行准确的辨认或确认，它具有语音获取方便、成本低廉、准确性高等特点。

近年来，计算机软硬件技术、通讯技术、网络技术、半导体技术和电子技术等的飞速发展不仅拓展了说话人识别技术的应用前景，也对其的发展提出了严峻的考验。如今，在安静的实验室环境下，说话人识别的正确率达到了 99%以上，然而在实际应用中，由于跨信道，背景噪声，声音的时变性，短语音等问题使得说话人识别技术的应用推广受到了严重的制约。

随着嵌入式时代的来临，智能设备逐渐终端化，移动化，小型化，随处可见嵌入式设备的身影，如 MP3、手机等。如今，人类越来越广泛的同智能设备进行交互，其交互形式多样化。但是依靠自然语言进行人机交互显然是最方便的。说话人识别系统从理论设想走向智能终端（如微型计算机、智能手机、其他嵌入式平台等）是发展的必然结果。该课题也是目前的一个研究热点，具有很高的商用

价值。

将说话人识别技术应用于嵌入式设备上有以下意义：说话人识别系统作为一种声纹加密工具，使用声音作为密码，安全简单可靠。将说话人识别应用于嵌入式平台，使得嵌入式产品能准确识别出主人身份，防止产品被没有授权的人使用，且方便移动作业，使得产品具有很好的市场竞争力。同时，嵌入式技术的发展反过来促进说话人识别技术的发展。随着嵌入式技术的发展，说话人识别技术也在不断地调整更新自己以便更好的适应嵌入式平台的发展。

总之，说话人识别系统应用于嵌入式设备上能进一步增强人们对说话人识别技术的关注，反过来，嵌入式设备上有了说话人识别系统的加入，功能更加多样化，安全性更高，具有更好的市场竞争力。

1.3 说话人识别发展及现状

说话人识别技术的发展经历了以下三个阶段：①技术启蒙阶段，即 20 世纪 30 年代，研究工作主要集中在人耳的听辨实验和探讨听音识别的可能性方面[2]。②技术突破阶段，即 20 世纪 60 到 70 年代，研究的重点为各种识别参数的提取、选择和实验上，并将倒谱比较和线性预测分析等线性处理和简单的模式匹配方法应用于说话人识别中[3]。③技术发展阶段，即从 20 世纪 70 年代末开始至今，说话人识别的研究转向对各种声学特征参数的非线性处理和新的模式匹配方法上[4]。

在说话人识别技术发展的三个阶段里，出现了许多具有标志性意义的成果。在理论方面，60 年代提出了动态规划(dynamic programming，DP)和线性预测分析技术（Linear Predictive Ceofficients，LPC），成为语音信号分析处理的强有力工具[5]；70 年代，线性预测技术进一步发展，动态时间弯折技术（Dynamic Time Warping，DTW）[6][7]基本成熟；80 年代，语音识别技术进一步深入，矢量量化（Vector Quantization，VQ）[8]和隐马尔科夫模型（Hidden Markov Model，HMM）[9]的提出标志着说话人识别技术的一个重大里程碑；此后人工神经网络（Artificial Neural Network，ANN）[10][11]，支持向量机（Support Vector Machine，SVM）[12]等理论不断被提出。

在应用方法，说话人识别技术已出现了一些比较成功的商用系统，50 年代，Bell 实验室实现了第一个可以识别是个英文数字的说话人识别系统——Audry 系统；80 年代至今，欧洲电信联盟开启完成的 CAVE 计划，实现了在电信与金融结合领域应用说话人识别技术、ITT 公司的 SpeakerKey 实现了电话声音的说话人确认等。在我国，由于汉语语音是一种声调语言，因此国外现有的一些技术成果无法直接使用。国内的说话人识别技术并没有特别广泛的商用性推广。目前国内较

成功的商用系统主要有：中科院自动化的 PATTEKSV 声纹识别和电话认证系统；科大讯飞语音实验室的 USTC-iF1 系统等。

1.4 说话人识别应用领域

随着时代的进步，说话人识别技术在国外已广泛的应用到诸多的领域，但是在国内，尚处于启动阶段，因此说话人识别技术在国内的发展空间更为广阔，在金融、证券、银行、公安、军队以及其他民用安全认证等行业和部门都有着广泛的需求[13]。目前，说话人识别主要应用于以下几个领域[14]：

民用领域：目前，常用的民用业务对用户的身份认证大多都是采用的数字密码，虽然方便简单但是安全性较低。在该领域将说话人识别技术同数字密码结合的方式可以更加安全有效地实现对用户身份的认证，且附加语音密码后还可以实现远程作业，这对用户来说更加安全便利。比如，电话服务中，以用户的声音完成查询、缴费等业务；用特定人的声音实现机密场所的出入人员检查，等等。

通信领域：在通信领域，说话人识别技术可以用于声音拨号、语音 E-mail、计算机远程登录、电话购物、信息服务、数据库访问、安全控制等。

司法领域：在司法领域，说话人识别技术可以对各种电话勒索、绑架等案件提供技术支持，可以根据录音查找出嫌疑人，帮助对嫌疑人的查证。

医学领域：说话人识别系统可以用于对特定患者的命令做出响应，如对假肢的控制等。

军事领域：说话人识别技术可以辨认出电话交谈过程中是否出现关键说话人，并对其交谈内容进行处理。另外，在对外发出军事指令时，可以实现对发出命令的人的身份进行确认。

1.5 说话人识别概述

1.5.1 说话人识别基本原理

说话人识别它同人类其他的生物特征（指纹、语音、虹膜等）一样，都具有普遍性，唯一性等特点。这些生物特征都能体现特定人与其他人的区别，且不容易被仿冒，可靠性高。鉴于说话人识别的研究对象是说话人的语音信号，且语音信号具有获取方便，成本低廉等特点，它比其他的生物特征更具有优势[15]。

说话人识别是指利用说话人语音中的能反应其独有的生理和行为特征的参数，来进行身份认证的一种技术。说话人识别技术分为训练（注册）和识别两个部分。所谓训练，是指对特定说话人的输入语音采取一系列的处理计算后提取能

表征说话人特点的特征信息，并对该特征进行建模的过程。识别，是指将待测说话人的语音特征同说话人模型进行比对，判断该语音是否对应为某个特定说话人[16]。图 1.1 所示为一个典型的说话人识别系统的框图。其由前端处理，模型训练，模式匹配和判决 4 个部分组成[13]。

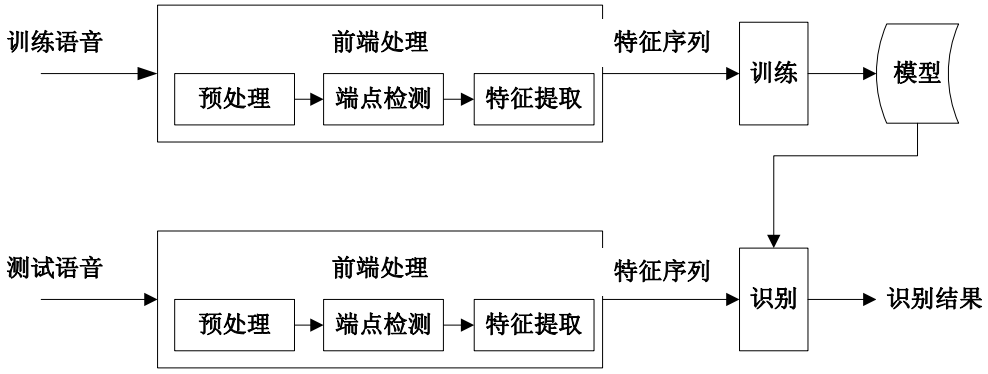

图 1.1 典型的说话人识别系统框图

由图 1.1 可以看出，无论是训练还是识别，都需要对输入的原始语音信号进行预处理，包括采用、量化、预加重、分帧和加窗等处理过程，以实现对语音信号进行特征提取。

1.5.2 说话人识别系统的分类

说话人识别系统依据不同的准则可以分为不同的种类[23]：

①根据识别语音的内容，可以分为文本无关（Text Independent）的说话人识别、文本相关（Text Dependent）的说话人识别和指定文本（Text-prompt）说话人识别。文本无关是指测试语音同训练语音的内容不需要相同；文本相关是指测试的语音同训练的语音内容必须相同；指定文本是指要求测试语音内容为系统指定的文本内容，不一定与训练语音内容完全一致。对一个文本无关的说话人识别系统而言，只要说话人相同，无需考虑测试文本内容是否同训练文本内容一致，系统也会予以“接受”。而对一个文本相关的说话人确认系统而言，在训练阶段，说话人需要对特定文本进行发音和建模，在识别阶段，只有说话人相同且发音的文本内容也相同，系统才会予以“接受”。由此可以看出文本无关的说话人识别系统相较与文本相关对用户更加友好，灵活性也更强，应用领域也更加广泛。虽然前者在实现难度上要大于后者，但是由于其具有很好的应用前景和实用价值，目前已经成为嵌入式开发的一个重要的研究领域。

②根据实际应用的范畴，可以分为说话人确认（Speaker Verification）和说话人辨认（Speaker Identification）。说话人确认指待识别的说话人语音只需同所申明的说话人模型进行匹配打分，最后由阈值来判定该测试语音是否通过，它的输出只有“接受”和“拒绝”两类，是一对一的问题；说话人辨认是指待识别的说话人语音同系统里所有说话人模型进行模型匹配打分，最后比较得分得出该语音是属于哪一个说话人，是多选一的问题。图 1.2、图 1.3 分别为说话人辨认和说话人确认系统框图。

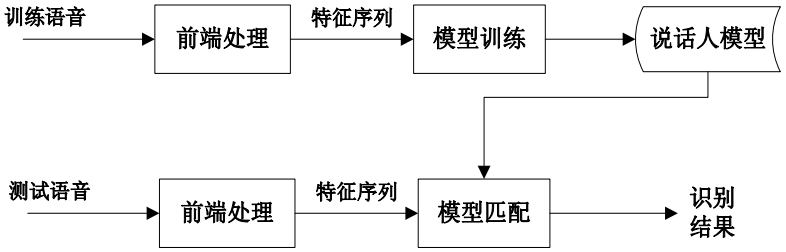

图 1.2 说话人辨认系统框图

由图 1.2 所示，说话人辨认系统是指对测试语音同模型库里的所有说话人模型分别进行比较，最终选择得分最高的作为识别结果。

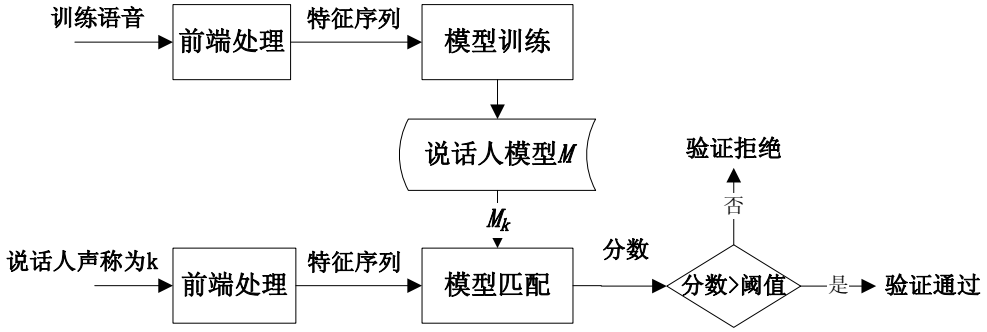

图 1.3 说话人确认系统框图

而图 1.3 所示，说话人确认系统则只是对测试语音同其所声称的说话人模型进比较，然后由阈值与得分进行比较得到识别结果。说话人辨认由于需要同各个模型进行逐个比对，因此计算量要远大于说话人确认。另外，说话人确认实现的是对说话人身份实现“接受”或者“拒绝”，安全性要高于说话人辨认，经济领域的应用也更广泛。基于以上两个因素考虑，本文将着重对文本无关的说话人确认技术进行研究。

③根据测试说话人身份的不同，可以分为开集（Open-set）的说话人识别和闭集（Close-set）的说话人识别。开集说话人识别要求测试的说话人集合没有训练集的限制；闭集说话人识别要求测试的说话人集合局限在训练集内的说话人内。

根据说话人识别的不同应用范畴，对说话人识别系统的评价标准也不一样。说话人辨认系统常用于对一段未知的语音信息，需要在大量的参考说话人中挑选出这段语音信息是属于参考说话人中的哪一个，常用的应用领域如刑事侦查中语音侦听。对于说话人辨认系统，其性能评价指标主要是正确识别率。即

$$RAT = N_{COR} / N_{tot} \tag{1.1}$$

其中 N_{COR} 为正确识别的个数，N_{tot} 为总测试的个数。

说话人确认系统中，使用者会声称自己是某一名说话人，由系统来判断其语音是否来自该声称的说话人，如身份认证、入境管理。对于说话人确认系统，其性能评价指标主要是等错误率（Equal Error Rate，EER），它包括两个参数：错误拒绝率（False Rejection Rate，FRR）以及错误接受率（False Acceptance Rate，FAR）。FRR 又称为Ⅰ型错误，它是将真实说话人误认为仿冒说话人所造成的错误，而 FAR 又称为Ⅱ型错误，它是把仿冒说话人误认为是真实说话人所引起的错误。EER 通常表示着两类错误均衡时的值，EER 越低表示系统的性能越好，在实际的应用场合，不仅需要考虑 EER 值，也需要考虑 FAR 和 FRR 的情况，这是因为不同场合对系统安全性要求不尽相同。

1.5.3 说话人识别技术难点

虽然目前说话人识别技术已取得了不少的成果，但是要达到成熟稳定的实际应用效果，仍有不少技术难点需待攻克[17][18]，主要表现为以下几个方面：

①说话人个性特征分离：语音信号里面往往包括了说话人的身份信息和说话内容信息，目前尚没有很好的方法将这两类信息从语音信号里进行分离。

②特征参数的自适应：说话人的发音常常与环境、情绪、健康状况等因素有关。如何使选取的特征参数具有自适性，目前尚没有找到特别有效的办法。

③复杂环境下的识别：目前说话人识别系统对环境的依赖性较强，如何在复杂环境下增强系统的鲁棒性也是说话人识别系统实用化前需要解决的一个技术难题。

目前的说话人识别系统大都是基于 PC 机的，但是随着嵌入式时代的来临，说话人识别系统从普通的 PC 机走向嵌入式平台也是大势所趋。相对于 PC 机嵌入式说话人识别系统面临着更大的机遇与挑战。其面临的技术难点主要表现为以下两个方面：

①有限运算存储资源下的运行效率。

嵌入式设备相对普通的 PC 机具有体积小，容易便携的优点，因此单纯地从能

耗上考虑，嵌入式设备是无法忍受过高的温度，因而在嵌入式设备上使用高频运算器并不合适，从而这也决定了其所具有的运算资源远远比不上普通的 PC 机器。说话人识别在实现中还包含了大量的复杂浮点运算，而目前大多的嵌入式设备都不具备浮点运算器。因此，嵌入式设备有限的运算存储资源和说话人识别巨大而复杂的运算处理之间的矛盾，是当前说话人识别应用于嵌入式平台的主要难点之一。

②更为复杂环境下的识别。

嵌入式设备便携的优点也决定了其所处的环境多变且复杂，同时嵌入式设备上的说话人识别系统还面临着跨信道、短语音、背景噪声等等之类的影响，这些对说话人识别系统的精度影响也是不容忽视的。如何在复杂的嵌入式使用环境中增强系统的鲁棒性也是如今说话人识别应用于嵌入式平台的主要难点之一。

1.6 论文研究工作和论文结构

1.6.1 研究思路和主要工作

目前，语音识别系统在嵌入式平台中已得到了广泛的应用[19]，如手机语音拨号，能识别主人命令的智能玩具，声控小车[20]等，但是将说话人识别技术应用于嵌入式设备上的应用却没有如此之广泛，目前对嵌入式说话人识别系统的研究大多集中在运算的优化，如定浮点转化运算，模型搜索空间压缩，复杂运算函数变形等，在算法层进行优化的并不多[21][22]。

由于嵌入式说话人识别具有良好的市场前景，越来越多的研究者投身于这方面工作。目前嵌入式设备上实现的说话人识别系统大多是基于特定人的文本相关的简单命令词识别，实用性差和灵活性低等特点制约了其在实际应用中的推广。基于以上的因素，本文工作重点定位于对文本无关的嵌入式说话人识别系统进行研究。

在上节中提到，嵌入式说话人识别系统实现的两大难点分别是：①嵌入式设备运算能力和存储资源有限同说话人识别系统复杂的运算之间的矛盾；②嵌入式设备使用环境不确定同说话人识别系统性能对环境依赖性强之间的矛盾。

为了使得文本无关的说话人识别系统在嵌入式设备上得以应用。我们从两个方面着手考虑：①提升系统的运行效率；②提升系统鲁棒性。

在提升系统的运行效率方面可以从以下两个方面来考虑：一是从嵌入式运行平台的具体情况入手，对程序中用到的复杂运算进行优化，降低单次运算的时间开销；二是优化程序算法结构，从根本上减少运算量来达到速度提升的效果。但

是，单纯的对运算过程进行优化往往意味着实现的简单化和粗糙化，必将对运算的精度带来损失，因而可能会对系统性能产生不利影响；而算法层面的改进，则有可能在加快运算速度的同时，也达到提升性能的效果。因此，本研究着重于优化算法结构来提升系统的运行效率。

在提升系统的鲁棒性方面可以从以下两个方面来考虑：一是去除语音中包含的噪音；二是寻找不受跨信道、背景噪声等因素影响的高层特征，将高层特征同低层特征进行结合来增强系统性能。但是对语音去噪本是一个很复杂的课题，而第二种方式相对简单，因此本研究着重对第二种方式进行研究与实现。

首先，汉语是一种声调语言，而声调在很大程度上反应了说话人的一种发音习惯，因此如果能够利用声调或者相关方面的特征进行辨识，就可以进一步提升系统的性能。

此外，说话人识别系统传统的做法多数是基于单一特征的辨识方法，如果我们可以结合不同种类的特征，如音段特征和超音段特征，则可以起到互相补充的作用，可以完整地对说话人的个性特征进行描述，得到更好的辨识效果。但是传统的基于高斯混合模型的方法不能够同时处理多种特征，因此需要考虑的问题还包括多种特征的结合方案。并且，由于各种特征是在不同层次或者是处于不同的方面，因此需要讨论在什么级别进行分类器的融合，以及特征空间的转换等一系列问题。

图 1.4 给出了本文的主要研究工作：

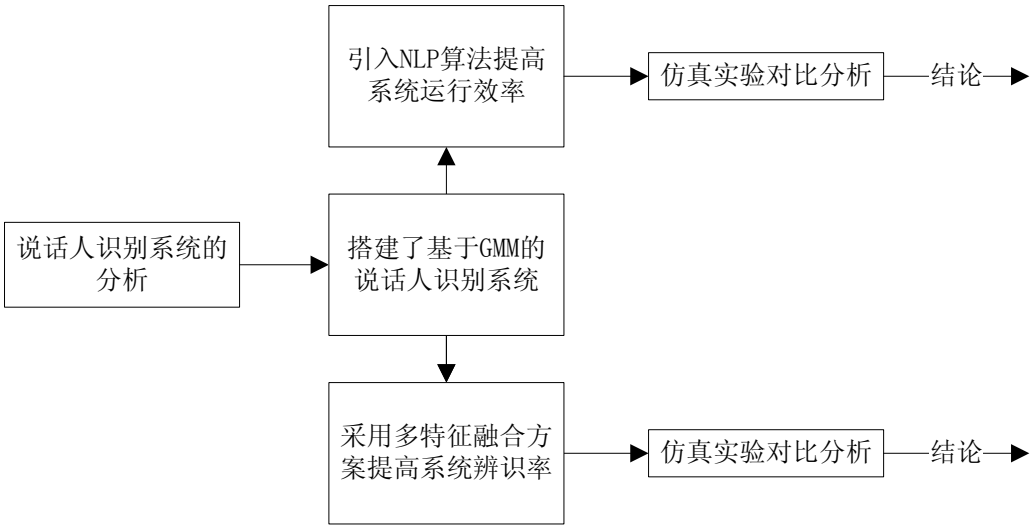

图 1.4 本文主要研究工作

本论文的改进和提出的新方法主要有以下两点：

①引入说话人识别领域里常用的非线性分段的方法来压缩信息量，提高程序的处理速度。并针对非线性分段方法中存在的不足，对非线性分段规则进行了改

进。

②采用了一种基于基频曲线段的特征。该特征不仅能够很好的模拟汉语声调的特点，还解决了韵律特征因为其不连续且长度不等而不能直接对其进行数学建模的问题。还提出了一种采用 SVM 对多重特征在分数层进行融合的方案。

1.6.2 本文的章节结构

本论文以下的各个章节将从不同的方面介绍本文的研究内容：

第 1 章，绪论。介绍了说话人识别技术的研究背景，阐述了目前该领域的研究现状和问题，以及本文的研究背景、意义和主要研究工作。

第 2 章，文本无关的说话人识别技术基础。主要对文本无关说话人识别系统的基本结构、各模块的处理方式和原理、系统的评价指标进行介绍。

第 3 章，语音数据库和基线系统设计。介绍了实验用的数据库信息以及基线系统设计的参数配置。通过实验验证基线系统的性能。

第 4 章，与文本无关的快速算法的研究。主要探讨了嵌入式平台上文本无关说话人识别的快速算法实现。介绍了非线性分段算法的思想和原理，并针对非线性分段算法在分段合理性方面的缺陷，提出了改进的算法提升识别性能。通过仿真实验验证算法的有效性。

第 5 章，与文本无关说话人识别多特征融合技术研究。探讨了采用多特征融合来增强系统鲁棒性的思路。主要内容包括采用了一种基于基频曲线的特征以及提出了一种新的采用支持向量机在分数层进行融合的方案。通过仿真实验验证方案的合理性。

第 6 章，总结与展望。对本论文工作进行了总结，并对下一步工作进行了展望。

第2章 文本无关的说话人识别技术基础

2.1 引言

说话人识别与语音识别的不同之处在于[23]：说话人识别是试图从语音信号中挖掘出能够表征人的个性特征，力图强调不同人之间的差别；而语音识别试图从语音信号中挖掘出表征语音内容的共同特征，力图对人的差别加以归一化。

在第 1 章中，我们通过对说话人识别技术的研究现状、研究意义和研究难点综述分析，将课题定位于面向嵌入式系统的文本无关的说话人识别技术的研究。本章将系统地对文本无关的说话人确认系统的各个子模块进行介绍。

2.2 说话人识别系统基本结构

说话人确认系统如图 2.1 所示，它是由前端处理、模型训练，模式匹配和判决等几大部分组成。说话人确认在判决阶段是跟预设的阈值进行比较，输出的结果只有接受和拒绝两种。

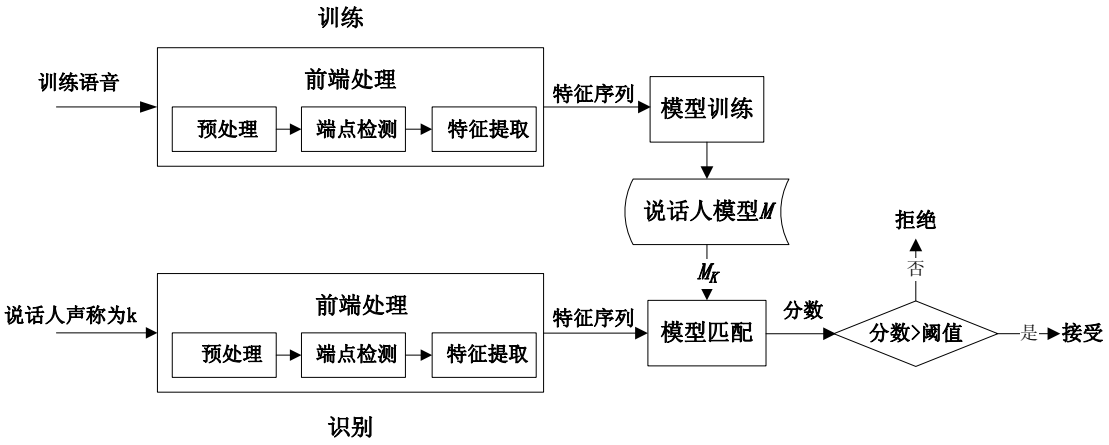

图 2.1 说话人确认系统框图

图 2.1 所示，说话人识别系统一般分为训练和识别两个部分。训练部分的主要步骤为：

①对训练语料进行分帧：按照设定的窗长和窗移对输入的语音进行分帧处理。

②前端处理：包括预处理，端点检测，特征提取。

③建模：将前端处理后的特征序列映射为一个模型。

识别部分的主要步骤为：

①对训练语料进行分帧：按照设定的窗长和窗移对输入的语音进行分帧处理。

②前端处理：包括预加重，端点检测，特征提取。

③模板匹配：将前端处理后的特征序列同模型库里的模型进行匹配计算，得到匹配分数。

④将匹配分数同预先设定的阈值进行比较来确定识别结果。

下面将对说话人识别系统中每一个步骤进行具体的介绍。

2.3 前端处理

2.3.1 采样和量化

从说话人的口中发出的语音在时间域和幅度域上都是连续的，而计算机的处理能力只限于离散数据，因此在采集语音的过程中，需要对信号进行离散化处理。在时间域上的离散化称为采样，而在幅度域上的离散化称为量化。

对一段语音进行处理的前提是语音信号数字化，数字化包括采样和量化。采样的方法一般是使用数模转换器（A/D），每隔一定长度的采样间隔进行一次采样[24]，采样间隔根据处理任务的不同而不同。采样间隔的倒数称为采样频率，例如在电话信道上的采样频率为 8 千赫兹（kHz），即每秒 8000 次，而在语音识别系统上的采样率一般为 16 千赫兹。

采样之后需要量化。常见的量化方法是在整个幅度值区间内等间隔地设定一些固定的离散点，称为量化电平，每一个语音采样之后的数据点被记录为离自己最近的电平值[23]。存储可以采用简单的二进制编码方案，即如果量化电平的个数为 16，则可以使用 4 位（bit）二进制来进行编码。这种方法称之为均匀量化，编码方法称为脉冲编码调制（Pulse Code Modulation，PCM）。

经过采样量化后的语音信号便可以进行前端处理了，前端处理包括：预处理、端点检测、特征提取三个部分。

2.3.2 预处理

预处理即是对语音信号进行高频加重，分帧加窗等操作。预加重实际上是在对语音信号进行采样和量化之后插入了一个高通滤波器，目的是为了提升高频信息，使得信号的频谱变得平坦，保持在低频到高频的整个频带中能够用同样信噪比求频谱，以便进行声道参数分析[23]。

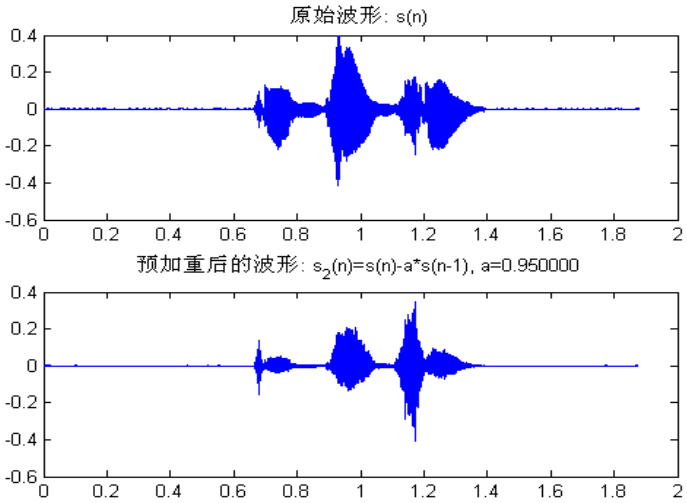

图 2.2 语音原始波形和经过预加重之后的波形

图 2.2 所示，语音内容是“中关村”，上图是没有经过预加重的原始波形图，下图是经过预处理之后的波形图。可以很明显的看出，经过预加重之后的高频信息被提升，但信号幅度却减弱了。

因为语音信号是非平稳、时变的，不能用处理平稳信号的数字信号处理技术对其进行分析。但通常认为在 10 毫秒到 30 毫秒的时间段中，语音信号是近似的平稳信号，几乎所有的语音信号的处理都是建立在“短时平稳”的基础上的。为了得到符合长度的短时语音信号，需要对语音进行加窗操作。窗函数（window function）将语音信号按照固定的长度分为帧（frame）,常用的窗函数有矩形窗，哈明窗（Hamming Window），哈明窗具有更平滑的低通特性，所以本文窗函数选取哈明窗，定义如下：

$$w(n)=(1-\alpha)-\alpha\cos\left(\frac{2\pi n}{N-1}\right),0\le n\le N-1 \quad , \qquad (2.1)$$

2.3.3 端点检测

端点检测目的是检测到语音信号的起始点和结束点。这样一方面减少了数据的存储空间，另一方面减少了后面的数据计算量。常用的端点检测法为帧能量检测法或过零率检测法[25]。

帧能量检测原理较为简单，实现也比较方便。它是使用帧能量同预估的阈值进行比较，帧能量高于阈值的，认为是语音帧，否则则视为非语音帧。在背景噪声较小的环境里，单一使用帧能量来进行端点检测能起到比较好的效果。

过零率（Zero Crossing Rate，ZCR）[26]描述的是时域波形穿过零线的频率。过零率区分清音和浊音时较为有效，清音的过零率较高，浊音过零率较低，无声部分的过零率基本为零。通常将帧能量同过零率联合使用能取得更好的效果，本系统采用是两者结合的方法进行端点检测。

对帧能量和过零率分别确定两个门限，一个是 T_L 代表较低的门限值，该值数值较小，对信号的变化比较敏感；另一个是较高的门限 T_H，数值较大。高门限被超过并且在自定义时间内的语音超过低门限，标记为信号的起始点。

此时整个端点检测可以分为四段：静音段、过渡段、语音段、结束段。静音段，当帧能量或过零率超过 T_L 时标记为起始点，进入过渡段。过渡段，当帧能量和过零率都低于 T_L 时，当前状态恢复到静音段状态。语音段，当帧能量或过零率超过 T_H 时，被认为是进入到语音段。处于语音段里的当前帧能量和过零率都降低到 T_L 以下，而且其总的计时时间少于最短的时间门限，则认为这段信号是一段噪音。接着继续侦测后面的语音信号，循环上面的操作，直到语音信号结束。

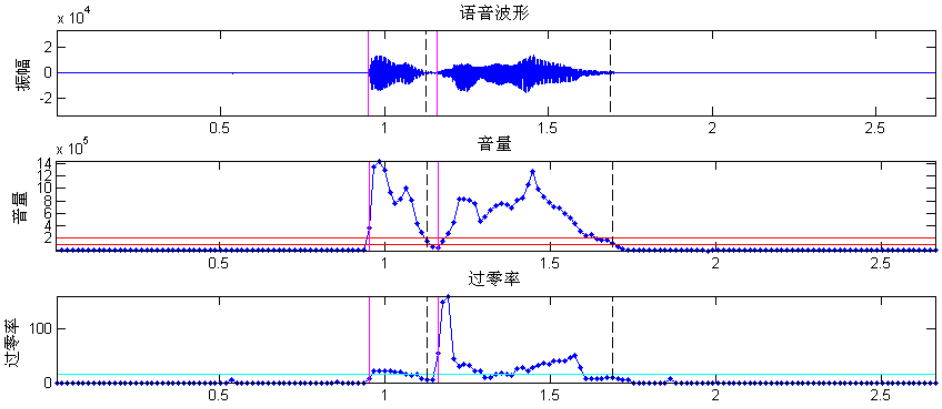

图 2.4 采用帧能量过零率端点检测仿真

图 2.4 中，实竖线表示语音段的起始位置，虚竖线表示语音的结束位置。由图可以看出，将帧能量同过零率联合使用进行语音起始点位置的检测，结果比较理想。

2.4 说话人特征参数提取技术

特征提取是说话人识别里的一项关键技术，目的是寻找一种能够合理地反映说话人个性统计模型的特征参数。所谓特征参数是指说话人自己本身的声学变异性（Inner Speaker Variation）尽可能小，而说话人同说话人之间的变异性（Inter Speaker Variation）尽可能大。

在理想的状况下，我们希望语音特征具有如下特点[27]：

①易于提取；

②能够有效区分不同的说话人；

③对同一说话人发音变化时能保持稳定性；

④不容易被模仿。

现在说话人识别系统里有良好表现并多被使用的特征是 LPCC 和 MFCC。线性预测是从与说话人个体密切联系的声道特征角度，描述了声道特性的全极点模型，因此通过线性预测分析得到的 LPCC 参数能够体现说话人的个人特征。而 Mel 频标是一种根据人耳的听觉特性设计的频标，通过对低频段语音细致地刻画，对高频稍微粗略的刻画，来模仿人耳对低频声音比对高频声音敏感的特性。

说话人识别任务中，MFCC 比 LPCC 具有更优越的识别性能，Thian 和 Kinnunen 等的研究结果显示，尽管存在各种各样的可选特征，然而 MFCC 的实际效果往往都是最出色的[28][29]。因此，本论文选择对说话人的 MFCC 参数进行建模。

2.4.1 梅尔频率倒谱系数

梅尔频率倒谱系数（Mel-scale Frequency Cepstral Coefficient，MFCC）,是基于人的听觉机理，考虑人耳对不同频率的感受程度，因此特别适合用于说话人识别里面，而且该特征参数对噪声环境也具有一定的鲁棒性，其差分倒谱能很好的反映语音的动态变化。梅尔频率能反应人耳对于频率 f（单位是赫兹）的感受度。他们的关系如下：

$$mel(f)=2595\times\log(1+f/700) \tag{2.2}$$

或是

$$mel(f)=1125\times\ln(1+f/700) \tag{2.3}$$

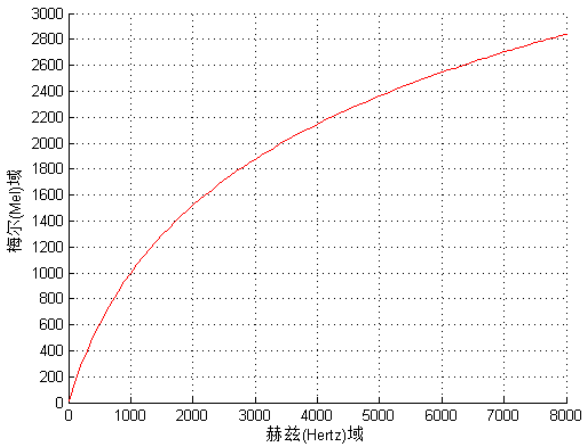

图 2.5　梅尔频率和赫兹频率的转换关系

梅尔频域和频率 f 的对应关系如图 2.5 所示，由此可以看出，人耳对于频率 f 的感受度是呈对数变化的。

MFCC 的完整计算流程见图 2.6 所示：

图 2.6　MFCC 提取流程图

具体的步骤如下：

①预处理：首先将语音信号通过一个高通滤波器，。这个目的是为了消除发音过程中声带和嘴唇的效应，来提升语音信号的高频部分。然后对预加重的语音信号进行分帧、加窗等处理后得到的时域信号。

②离散傅里叶变换：将第 1 步得到的加窗后的信号经过 FFT 变换得到频率谱上的能量分布。这是因为信号在时域上的变化很难看出它的特性，所以通常要将他们转换成频率上的能量分布来观察，不同的能量分布代表不同的语音特征。

③构建三角滤波器组：需要注意的是，这些三角滤波器组在梅尔频率上是均匀分布的，在低频部分，人耳的感受比较敏锐，在高频部分，人耳的感受会越来越粗糙，如图 2.7 所示：

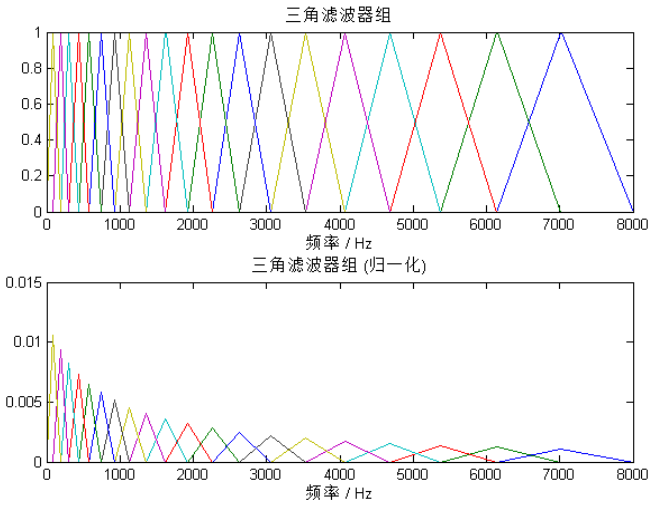

图 2.7　梅尔滤波器组

将能量谱能量通过梅尔滤波器组得到对数能量，该步骤是为了对频谱进行平滑，消除滤波的作用，突显原语音的共振峰，也就是说，MFCC 参数内是不包含该语音的 pitch 值的。

④将经过离散余弦变换（Discrete Cosine Transform，DCT），求出 L 阶的梅尔频率倒谱系数，在说话人识别技术中，L 通常取 16。离散余弦的变换公式如下：

$$c(n)=\sqrt{\frac{2}{N}}\sum_{m=1}^{N}F(m)\cos(\frac{\pi n}{N}(m-0.5)),n=1,2,...,L \tag{2.4}$$

其中，N 表示三角滤波器的个数。这样求出的 MFCC 参数的维数为 L。

⑤动态参数。对于一段语音来说，相邻帧之间的语音信息是相关联的，但是梅尔频率倒谱系数是以独立帧为单位进行计算的，为了反映帧间的相关性，常需要将动态参数补充到原静态特征向量里。计算公式如 2.5 所示：

$$\Delta^{m}c(n)=\Delta c(n+m)-\Delta c(n-m) \tag{2.5}$$

其中 m 表示阶数，一般取 1 或 2，表示静态参数。于是加上了一阶动态参数的特征矢量表示为，若取 L 为 16，则加上一阶差分系数后的特征矢量变成了 32 维。

2.4.2 基于基频曲线多项式拟合

声音信号泛指人耳听到的各种声音信息，它是由发声体发生振动，对空气造成压缩或者扩张形成声波，声波经由空气传送至人耳，耳膜感受到声波的振动，再由神经系统将振动信号传送至大脑，最后由大脑进行解读此声音信号的意义。声音信号根据不同的分类标准有多种分类方式：

①发声体的种类分类。可以分为生物声音信号（如人声、狗声等）和非生物声音信号（雷声、雨声、开关门声等）。

②声音信号的周期性分类。可以分为准周期音和非周期音。准周期音是指声音的波形具有规律性，人耳可以感觉其存在稳定音高，如人的歌声等；非周期音是指声音的波形不具有规律性，人耳无法感觉其稳定音高的存在，如打雷声等。

(1) 基频曲线

说话人识别技术探讨的主体是人声。人声是由浊音（Voiced sound）和清音（Unvoiced sound）两部分组成。浊音是由声带的振动所发出的声音，具有规律性；清音是由嘴唇发出的气音，没有规律性。浊音信号具有较强的周期性，称为基音周期。基音周期的倒数称为基音频率，也称为 pitch 或者 f0，基频反映的是声带振动的快慢，振动越快，声音频率的越高。而清音没有周期性，故清音没有基频信息。

基频信息是语音信息中的重要参数，它在旋律辨识，语音合成等领域都有广泛的应用。有研究证明考虑基频信息可以增加语音识别系统的性能[30]。

图 2.8 和图 2.9 分别是用 praat 软件提取出来的 A,B 两人说同一句话的波形图

以及基频曲线图。语音内容是“中关村”，图的上方是语音波形图，下方是语谱图，图中的点是基频数据点，由点连接的曲线是基频曲线。

由图可以看出，一句话的基频曲线段并不是一条连续的曲线，而是由几段不连续片段构成的，这也证明了只有浊音帧才有基频信息。另外也可以看出，虽然两个人用普通话朗读同样的内容，但是两个人的基频曲线段却不太一样，可以认为基频曲线的超音段特征包含了说话人区别于他人的个性信息，在说话人识别系统中考虑基频信息能够增强系统的性能。

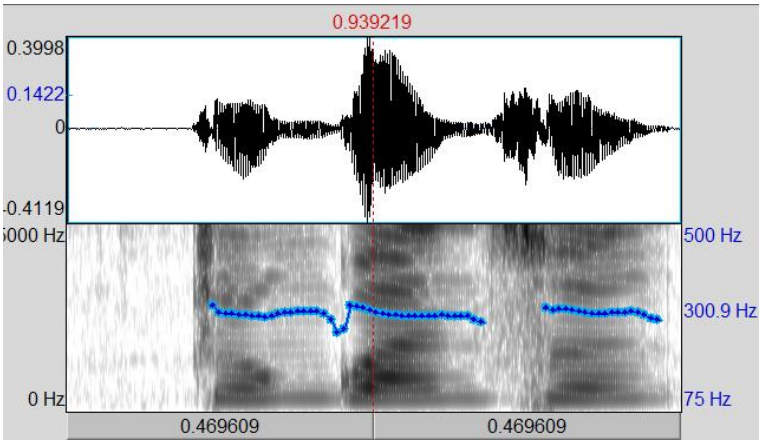

图 2.8 说话人 A 的波形，基频曲线段图

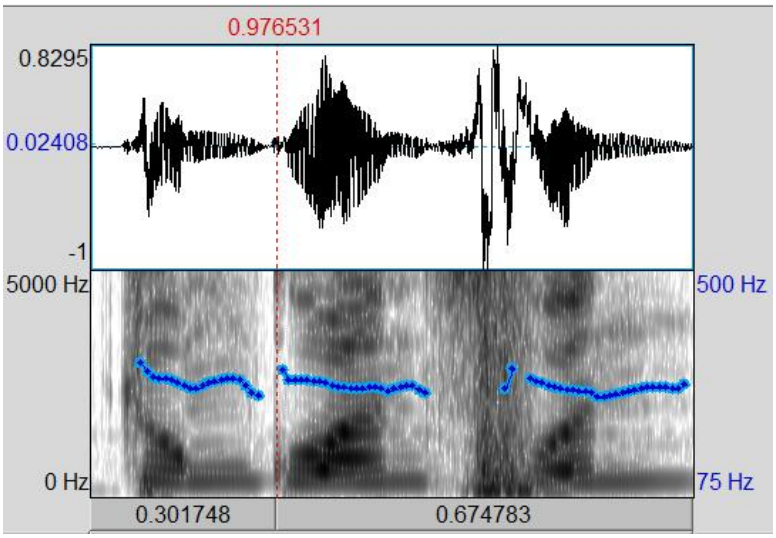

图 2.9 说话人 B 的波形，基频曲线段图

基频信息的提取流程如下：

①对整段语音信号进行分帧，帧与帧之间可以重叠；

②计算每帧的 pitch 值（常用 ACF（Autocorrelation function）或 AMDF（Average

magnitude difference function））；

③排除不稳定的 pitch 值。（常用 pitch 值的范围来筛选）；

④对整段 pitch 值进行平滑处理（常采用中通滤波器（Median Filters））。

(2) 多项式拟合基频曲线特征

一句话中每段基频曲线长度不一定相等，在采用统计的方法对一个人的基频信息进行建模时会遇到样本点维数不等的情况而无法进行数学建模。要将基频信息用在文本无关的说话人识别系统里需要对基频信息作一些处理。

我们考虑用多项式拟合的方法来描述每段基频曲线的形状特征，采用多项式的系数作为处理后的基频曲线段特征。多项式的系数具有物理意义，比如抛物线的系数代表了抛物线的开口大小和方向。这样用多项式系数构成的特征不仅具有固定长度，还能反应基频曲线段的高度，弯折信息。采用的具体的拟合算法是最小二乘法。

最小二乘法的描述如下[31]：

对于每一段基频曲线，我们用 $\{(x_i, y_i)\}$（其中 $i = 0,1,...,m-1$）来表示每一个基频的数据点。其中，m 表示数据点的总个数，x_i 表示该点出现的相对时间，y_i 表示该点具体的基频数值。我们需要求的是一个函数 $y = S^*(x)$，每个数据点上的误差就是 $\delta_i = S^*(x) - y_i$。

记总的误差为一个向量 $\delta = (\delta_0, \delta_1, ..., \delta_{m-1})^T$。目标是让误差的平方和最小化，也就是说要在线性无关函数族 $\psi = span\{1, x, x^2, ..., x^n\}$ 上找到一个函数 $S^*(x)$，使得

$$\|\delta\|_2^2 = \sum_{i=0}^{m-1} \delta_i^2 = \sum_{i=0}^{m-1} [S^*(x_i) - y_i]^2 = \min_{S(x)\in\psi} \sum_{i=0}^{m-1} [S(x_i) - y_i]^2 \tag{2.6}$$

此时

$$S(x) = \sum_{j=0}^{n} a_j x^j = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n \tag{2.7}$$

于是问题转化为一个求多元函数

$$I(a_0, a_1, ..., a_n) = \sum_{i=0}^{m-1} (\sum_{j=0}^{n} a_j x_i^j - y_i)^2 \tag{2.8}$$

极小值的问题，这里函数的变量是 $a_0, a_1, ..., a_n$。

由多元函数极值的必要条件，有

$$\frac{\partial I}{\partial a_k} = 2\sum_{i=0}^{m-1} (\sum_{j=0}^{n} a_j x_i^j - y_i) x_i^k = 0 \tag{2.9}$$

展开并移项，得

$$\sum_{j=0}^{n}(\sum_{i=0}^{m-1}x_i^{j+k}a_j)=\sum_{i=0}^{m}x_i^k y_i \tag{2.10}$$

写成矩阵形式，就是

$$\begin{pmatrix} \sum x_i^o & \sum x_i^1 & \cdots & \sum x_i^n \\ \sum x_i^1 & \sum x_i^2 & \cdots & \sum x_i^{n+1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ \sum x_i^n & \sum x_i^{n+1} & \cdots & \sum x_i^{2n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{pmatrix} = \begin{pmatrix} \sum x_i^0 y_i \\ \sum x_i^1 y_i \\ \cdot \\ \cdot \\ \cdot \\ \sum x_i^n y_i \end{pmatrix} \tag{2.11}$$

这样，只需要利用已知的所有(x_i, y_i)来求解上面的矩阵方程组（称为法矩阵），就可以得到合适的多项式系数$(a_0, a_1, ..., a_n)$。

经过这样的处理之后，每一个基频曲线片段都转化成了一个长度（维数）为的特征向量,将这些系数向量组成能被系统利用的基频曲线段特征。

(3) 多项式阶数选定

另外，需要注意的是多项式的阶数设定并不是越大越好，阶数设定还需根据实际情况而定。观察基频曲线片段，我们发现基频曲线的顶点个数都在三个以下，因此多项式阶数设定为 3，图 2.10、图 2.11 分别显示的是图 2.8、图 2.9 中两个人基频曲线作为多项式拟合的例子。

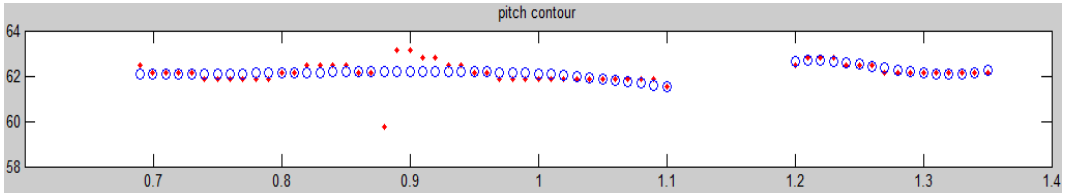

图 2.10　说话人 A 原始基频曲线和拟合后的基频曲线

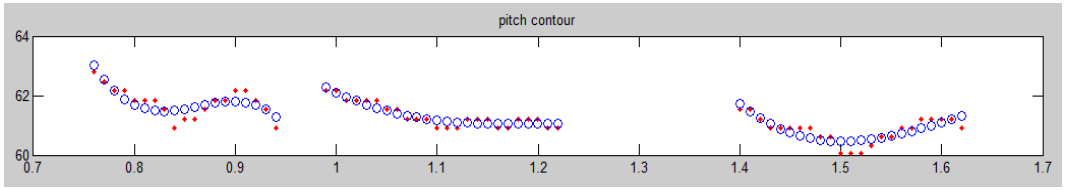

图 2.11　说话人 B 原始基频曲线和拟合后的基频曲线

图 2.10、图 2.11 中，实心点表示原始的基频曲线数据点，圆圈表示采用三次多项式拟合后的数据点所在的位置。由两图可以看出，采用多项式拟合的方法得到特征基本可以代表原始基频曲线的主要特征，如高度，开口方向，弯折程度等。

另外，将多项式的阶数设置为 3 也较为合适。所以本文后续的实验用到该方法都是将阶数固定为 3。

2.5 说话人建模方法

2.5.1 矢量量化

矢量量化(Vector Quantization, VQ)[8][32]是一种非参数模型方法，它常用 K-均值算法和 LBG 算法作为匹配算法。在训练阶段，通过对说话人语音提取能够表征说话人身份的特征，采用 VQ 技术将这些特征矢量浓缩成具有代表性的码本，将这些码本作为说话人的模板，在识别阶段用和训练阶段相同的方法提取测试模板，通过匹配比较测试模板同说话人模板间的距离，以距离最小的识别结果。矢量量化 VQ 技术不依赖于参数的时间顺序，相比常用于文本相关的动态时间弯折（Dynamic Time Wrapping, DTW）更为简单和快速。但是该方法对信号和背景噪声的依赖性比较严重，这两种影响可以改变说话人特征空间的码本的位置，最终导致模板漂移。

2.5.2 隐含马尔科夫模型

隐含马尔科夫模型(Hidden Markov Model，HMM)可以用在与文本有关或者与文本无关中。在文本相关的说话人识别中，由于说话人特征矢量构造时间是确定的，所以常采用从左至右的 HMM 或 Bakis HMM，但是在文本无关的说话人识别系统中，常采用各态经历 HMM。训练时，先对说话人的特征矢量建立其文本无关的各态经历 HMM，采用 Baum-Welch 估算 HMM 参数作为特征模板，测试时，将测试语音同参考模型进行匹配计算，得到识别结果。训练常用的算法是 Baum-Welch 算法，识别常用的算法是 Viterbi 算法。

2.5.3 高斯混合模型

高斯混合模型(Gaussian Mixture Model，GMM)[33]由美国学者 Douglas A. Reynolds 于 1995 年引入说话人识别领域，是一种通用的概率模型，能有效的模拟多维矢量的任意连续概率分布，因而很适合在文本无关的说话人识别中为说话人建模[34]。

GMM 可以看成是单状态的混合连续分布 HMM，它采用多个高斯分布的线性组合来近似说话人的特征分布。训练时，采用多个高斯函数对说话人的特征矢量进行聚类，每一个高斯函数的均值，协方差矩阵和出现的概率代表一类，将这些

类的参数组合起来形成说话人参考模板。识别时，将测试语音同参考模板进行匹配计算似然分，以似然分最大的作为识别结果。GMM 的训练是采用 EM（Expectation-Maximization）算法。

基于高斯函数的线性组合具有描述各种状态不同分布的能力，在高斯数合理的情况下，GMM 能够给出对任意的函数分布有效逼近。基于这个理论，本论文的研究工作都是基于 GMM 进行的。

(1) 高斯混合模型定义

高斯混合模型是用高斯概率密度函数（Probability Density Function，PDF）对事物进行精确量化，将其分解为若干个基于高斯概率密度函数的分布而形成的模型。它由若干个概率密度函数及其相应的权重组成[35]。将高斯混合模型应用到说话人识别中，做法是将说话人的声学特征进行分类，然后每一类特征由一个高斯密度函数进行描述。如图 2.12 所示。

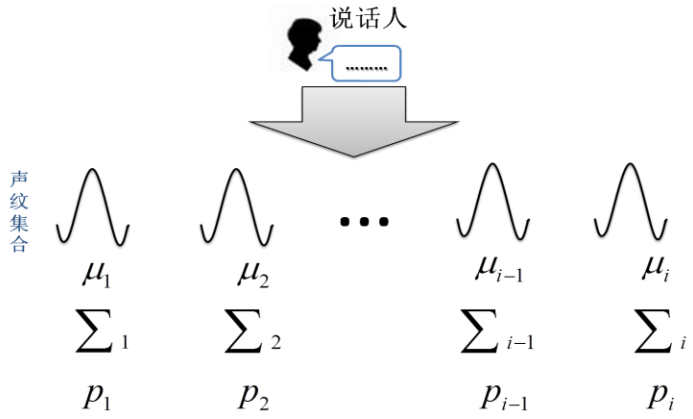

图 2.12 高斯混合模型

多元正态分布的概率密度函数是：

$$p(x)=\frac{1}{(2\pi)^{\frac{d}{2}}|\sum|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(x-\mu)^T\sum{}^{-1}(x-\mu)\}\equiv N(x|\mu,\sigma) \tag{2.12}$$

其中，x 是一个 d 维随机向量，μ 是概率分布的均值向量，$\sum$ 是一个 d 行 d 列的协方差矩阵。由于语音特征参数的各个维度之间是相对独立的，因此协方差矩阵 $\sum$ 就退化成为对角矩阵。这在很大程度上简化了模型，并且节省了训练和测试时需要的计算量和时间。

高斯混合模型是一组带权重的多元概率密度函数，多元概率密度函数的个数称为高斯混合模型的混合数，记为 m；各个概率密度函数的权重（weight）记为 $\{\omega_j\}$；每一个概率密度函数称为一个高斯分量（Gaussian component），记为 $\{p(x|j)\}$。

在进行概率的计算时，实际上是多个概率密度函数的加权求和，也就是：

$$p(x)=\sum^{m}\omega_j p(x\mid j) \tag{2.13}$$

参数需要满足的归一化限制条件是： $\sum^{m}\omega_j=1, j=1,2,...,N$ ， $0<\omega_j<1$ ，

$\int p(x\mid j)dx=1$ 。

(2) 高斯混合模型的训练：EM 算法

EM 算法[36][37]主要由 E 步骤（估计期望值）和 M 步骤（重估参数）组成。EM 算法类似一个逼近算法，通过计算每个训练样本在选定的初始参数下的概率估计，不断地修正模型参数。这样不停地迭代计算，直到满足收敛条件才终止。

对于高斯混合模型来说，参数的初始化有多种方法，可以随机指定各个高斯分量的均值点，各个分量的权重设为相同，协方差矩阵设为单位矩阵。但是通常来说，EM 算法可能需要较多次数的迭代才能够收敛，所以在初始化阶段可以采用更加简单而方便的方法来获得一个相对可靠的初始值。通常采用 K-means 算法进行聚类，随后在聚类的基础上开始迭代计算。

K-means 流程简要介绍如下：

①从所有训练样本中，随机选取k个作为初始聚类中心；

②对于所有的训练样本，计算其到每一个聚类中心的距离，并选择距离最近的一个聚类中心，并认为该样本属于该类；

③对于每一个聚类，计算其内部所有元素的均值，作为新的聚类中心；

④如果聚类中心有变化，则返回第 2 步，否则结束。

在进行初始化之后，我们将所有的样本点聚类得到了k个初始类。各个类的中心作为各个高斯分量的初始均值，各个类内部各维参数的方差构成对角协方差矩阵，各个初始聚类内部的样本数量作为各个高斯分量的权重。然后开始用 EM 算法进行迭代。

EM 算法[36][38]的推导和计算需要引入隐藏变量来描述样本属于哪一个高斯混合分量，同时需要计算隐藏变量的数学期望。这里只给出最终的训练的流程：

①初始化均值 μ_k、协方差矩阵 $\sum_k$、混合系数 ω_k、并计算初始的对数似然分。

②E 阶段。用现有的参数计算可靠性：

$$\gamma(z_{nk})=\frac{\omega_k N(x_n\mid\mu_k,\sum_k)}{\sum_{j=1}^{m}\pi_j N(x_n\mid\mu_j,\sum_j)} \tag{2.14}$$

③M 阶段。用现有的可靠性结论来重新估计参数：

$$\mu_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n \tag{2.15}$$

$$\sum{}_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k^{new})(x_n-\mu_k^{new})^T \tag{2.16}$$

$$\omega_k^{new}=\frac{N_k}{N} \tag{2.17}$$

这里 $N_k=\sum_{n=1}^{N}\gamma(z_{nk})$ 。

重新评价对数似然分：

$$\ln p(X\mid\mu,\sum,\pi)=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{m}\omega_k N(x_n\mid\mu_k,\sum_k)\} \tag{2.18}$$

然后检查是否符合收敛准则，如果不符合则转回到第 2 步。

(3) 高斯混合模型的测试

在训练阶段，我们对每一个目标说话人的语音建立一个高斯混合模型。

在测试阶段，简单地说是将测试语音在每一个高斯混合模型上进行似然分的计算，随后选择得分最大。但是和之前的描述有所不同的是，由于输入的每一个样本都含有多个语音帧，因此在一个样本的每个帧上都需要进行似然分的计算。对于整段语音样本而言，我们认为各个帧的特征之间是彼此独立的，因此在计算每个帧的得分之后，需要用求乘积的方法获得所有帧在每一个模型上的得分。

测试阶段的算法流程描述如下：

设输入的待辨识样本 x 共由 k 帧组成，记为 $x=(x_1,x_2,...,x_k)$。那么第 t 帧数据在第 j 个高斯混合模型 g_j 上的似然分可以用下面的公式进行计算：

$$p(x_t|g_j)=\sum_{i=1}^{m_j}w_{j,i}p_{j,i}(x_t|g_j) \tag{2.19}$$

其中， $p_{j,i}(x_t|g_j)$ 是 x_t 在高斯混合模型 g_j 的第 i 个分量上的似然得分：

$$p_{j,i}(x_t|g_j)=\frac{1}{(2\pi)^{d/2}\left|\sum j,i\right|^{1/2}}\exp\left\{-\frac{1}{2}(x_t-\mu_{j,i})^T\sum{}_{j,i}^{-1}(x_t-\mu_{j,i})\right\} \tag{2.20}$$

因此，整个待辨识语音x所含有的每一帧的似然得分的乘积构成了x关于模型g_j的得分，也就是

$$p_{j,i}\left(x_t|g_j\right)=\prod_{t=1}^{k}p\left(x_t|g_j\right) \tag{2.21}$$

上式中包含全部k帧得分的连乘积，为了避免可能出现的下溢问题而影响精度，在实际应用中，常常使用计算对数似然分（log likelihood score）然后再求和的方法，即

$$\log p\left(x|g_j\right)=\log\left(\prod_{t=1}^{k}p\left(x_t|g_j\right)\right)=\sum_{t=1}^{k}\log p\left(x_t|g_j\right) \tag{2.22}$$

2.5.4 人工神经网络

人工神经网络(Artificial Neural Network，ANN)在一定程度上可以模仿人脑的功能，是一种模仿生物神经网络的结构和功能的数学模型或计算模型。说话人识别使用过多种类型的神经网络类型。前向神经网络（BP）一期结构简单、分类性能好在说话人识别中得到了广泛的应用；多层前向神经网络（RBF）是映射神经网络，它可完成从说话人特征空间向说话人集合的映射。采用神经网络进行说话人识别，如果使用一个网络作为分类器，当待识别的人群（N）改变时，网络的结构（至少输出神经元个数）也将随之发生改变，需要对网络进行重新训练。另外，如果 N 增大时，神经网络的训练时间将以指数增大，目前解决办法是将当个大网络分解成许多部分功能的子网络，再将子网络进行组合来完成大网络的功能。

2.5.5 支持向量机

支持向量机方法[39]是根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力。支持向量机方法的几个主要优点如下：①专门针对有限样本集，可以再现有的信息下得到最优解，解决了神经网络中无法避免的局部极值问题；②它将实际问题通过非线性变换转换到高维特征空间，在高维特征空间构造线性判别函数来实现原空间中的非线性判别函数，其算法复杂度和样本的维数无关，不仅巧妙地解决了维数问题也使得机器有较好的推广能力。如今，SVM 已经成功地应用到模式识别领域中，并表现出了良好的性能，成为研究的新热点。因此，本文研究工作将支持向量机作为多特征融合系统里的分类器融合工具。

(1) 支持向量机定义

一般来说，我们的模式分类或辨识系统的流程是首先对训练样本进行建模，然后对测试样本进行分类辨识。如果把样本看成是高维空间的点，把样本之间的分界面看成是高维空间的平面或者曲面，那么所谓的分类辨识，就是看测试用的样本点被哪些曲面所包围，或者说在曲面的哪一侧。

支持向量机的主要思想可以概括为两点：①它是针对线性可分的情况进行分析，对于线性不可分情况，通过使用非线性映射算法将低维输入控件线性不可分的样本转化为高维特征空间使其线性可分。从而使高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。②它基于结构风险最小理论之上的特征空间中构建最优分割超平面，使得学习器得到全局最优化，并且在整个样本空间的期望风险以某个概率满足一定的上界。

考虑一个在二维平面上的两分类辨识问题，即所有的样本点都处于二维平面上，且所有的样本点只有两类，如图 2.13 所示。

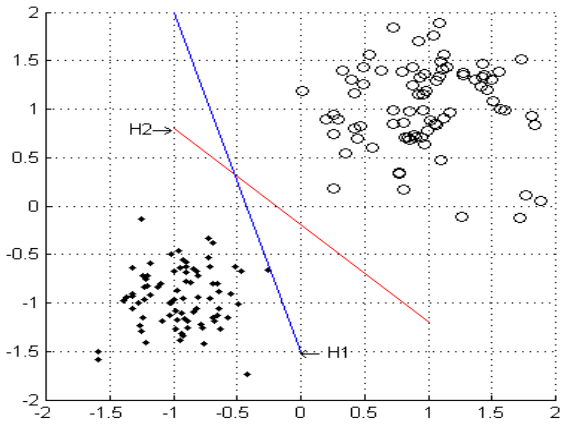

图 2.13 SVM 对数据分类

显而易见，图 2.13 中的点（分别用实心点和空心圆形表示两个类别）可以用一条直线（实际上，是二维空间上的“平面”）来进行区分，并且这样的分割线（面）不止有一个，图中的直线 H_1 或者 H_2 都可以进行分隔。但是从另一个角度，我们总是认为直线 H_2 比直线 H_1 要“更优”，原因就是看起来它“可以更好地将两个类别进行划分”。换句话说，对于和训练样本同分布的测试样本来说，“越过”直线 H_1 而导致分类辨识错误的可能性要大于越过 H_2 的可能性。原因就是，直线 H_2 不但进行了样本的分隔，同时还使得两类不同样本之间的间隔最大，进而在测试集上就更有可能获得最佳的分类辨识效果。对于这样的直线 H_2，我们称其为最优分界线，而在一般的高维空间里面，称为“最优分界面”，同时，这样的分类器称为“最大

间隔分类器”。

在低维空间中，样本不一定是线性可分的，或者说用线性分界线（面）进行分类时会有较大的误差，但是如果应用某种特征空间的变换方法，使得特征点被映射到了高维空间中，就有可能进行线性分隔或者获得更低的辨识错误率。

考虑图 2.14 所示的例子。左图中实心点和 x 形标记分别表示在二维平面上的两种不同类别的样本点。实心点近似分布在一个圆周上，而 x 形点则近似分布在中央的正方形区域内。显然，在二维平面上这两类点是无法通过直线进行分隔的，并且通过线性变换也难以完成。但是，如果将这些样本点映射到三维空间，其 z 坐标按照公式 $z_i = \sqrt{x_i^2 + y_i^2}$ 计算，即每一个点的 z 坐标，等于它在原始二维平面上与原点之间的距离，那么在三维空间中，这些点的位置如图 2.14 右边的一幅图所示。显然它们可以用图中的平面分隔开来。

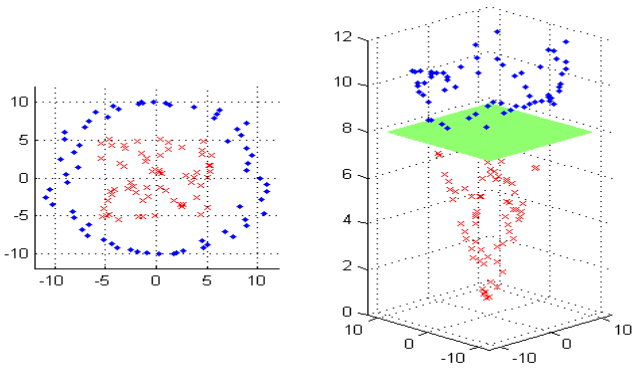

图 2.14 映射到高维空间中，数据点变成线性可分的了

支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面。分隔超平面使两个平行超平面之间的距离最大化。那么平行超平面间的距离越大，分类器的总误差就越小。如图 2.15 所示。

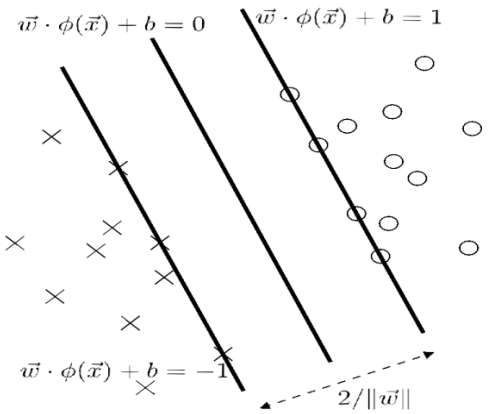

图 2.15 两类样本用 SVM 训练得到的最大间隔超平面

在支持向量机的训练过程中，我们的目的是求出这样的距离最大的两个超平面。这样的超平面上面的数据点被称为支持向量（support vector），因为这些向量直接“支持”起了所需要的超平面，支持向量机的名字也是由此而来的。

(2) 支持向量机的训练以及参数选择

对于一般的样本，我们需要寻求一种合适的非线性映射方法，把样本映射到比原来特征空间的维数高得多的映射空间。因此，支持向量机的“训练”过程，就转化成为两个基本问题：如何在高维空间寻找最优分界面，以及如何寻找合适的非线性映射函数。

我们从较为简单的二分类问题入手，来描述支持向量机的训练方法。

记现有的数据点集是$(\vec{x}_i, y_i), i=1,2,\ldots,n$。其中$\vec{x}_i$是$d$维向量，$y_i$是标注，取值为$\{1,-1\}$，分别表示数据点的两种不同类别。

在二维平面上，一般直线的方程可以写为$y=kx+b$，这里k和b都是实数（垂直于x轴的特殊直线可以通过坐标轴的旋转转化为普通直线，因此暂不考虑此种情况）。如果这样的直线是一个分界面，那么要判断某一个样本点(x_1, y_1)是在直线的哪一边，只需要计算判别式$f(x,y)=f(x_1,y_1)=kx_1-y_1+b$的符号即可。

推广到一般的高维空间和对应的高维平面，用$\phi(\vec{x})$表示一种固定的特征空间映射，判别式可以写为

$$f(\vec{x})=\vec{w}\cdot\phi(\vec{x})+b \tag{2.23}$$

这里定义判别式符号为正的时候是第一类，符号为负的时候是第二类。

我们可以对分类面函数进行归一化，使得所有的点都满足$y_n(\vec{w}\cdot\phi(\vec{x})+b)\geq 1$，等号对于那些恰在分类面上的点成立（如图 2.15 所示）。这样，两个类之间的间距就是$2/\|\vec{w}\|$。要让它最大，就是让它的倒数最小，因此求最优分类面的问题就转化成了求式(2.24)的最小值问题。

$$\arg\min_{\vec{w},b}\frac{1}{2}\|\vec{w}\|^2 \tag{2.24}$$

定义拉格朗日（Lagrange）函数如下：

$$L(\vec{w},b,\vec{a})=\frac{1}{2}\|\vec{w}\|^2-\sum_{n=1}^{N}a_n\left\{y_n\left[\vec{w}\cdot\phi(\vec{x})+b\right]-1\right\} \tag{2.25}$$

其中$\vec{a}=a_1,\ldots,a_N$是拉格朗日系数。

多项式极值的条件是对于各个变元的偏微分为 0，也就是

$$\frac{\partial L(\vec{w},b,\vec{a})}{\partial b}=0 \text{ 且 } \frac{\partial L(\vec{w},b,\vec{a})}{\partial \vec{w}}=0 \tag{2.26}$$

即

$$\sum_{i=1}^{N} a_i y_i = 0 \tag{2.27}$$

$$\sum_{i=1}^{N} a_i \phi(\vec{x}) y_i = \vec{w} \tag{2.28}$$

代入原拉格朗日函数，转化为对偶问题：

$$\tilde{L}(\vec{a}) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j \phi(\vec{x}_i)\cdot\phi(\vec{x}_j) \tag{2.29}$$

这是一个不等式约束下的二次函数极值的问题。约束条件就是 $a_i \geq 0$，$\sum_{n=1}^{N} a_n y_n = 0$。求解之后可以得到唯一解，记为 $\vec{a}^*$，此时

$$\vec{w}^* = \sum_{i=1}^{N} \vec{a}^* y_i \phi(\vec{x}_i) \tag{2.30}$$

由 Kuhn-Tucker 条件，这个优化问题的解需满足：

$$a_i\left(y_i\left(\vec{w}\cdot\phi(\vec{x}_i)+b\right)-1\right)=0, i=1,...,n \tag{2.31}$$

因此，多数的 a_i^* 将为 0，取值不为零的 a_i^* 对应于支持向量，于是可以解出 b^*。最后，得到的最优分类函数是

$$f(\vec{x}_i)=\text{sgn}\left\{\vec{w}^*\cdot\phi(\vec{x}_i)+b^*\right\}=\text{sgn}\left(\sum_{i=1}^{N} a_i^* y_i\left(\phi(\vec{x}_i)\phi(\vec{x})\right)+b^*\right) \tag{2.32}$$

(3) 非线性可分情况的讨论及和函数的选择

前面的方法只考虑了所有样本都是线性可分的情况。在实际问题中，通常遇到的情况都是线性不可分的，即不存在这样的分界面可以把全部训练样本根据其标注进行完全正确的划分。

解决方法是，我们在判别式中，增加一个松弛项，使得所有的点都满足：

$$y_n\left(\vec{w}\cdot\phi(x_n)+b\right)+\xi_i \geq 1 \tag{2.33}$$

那么最优分类面问题仍然是求最小值问题，但是增加了参数，这里的参数 C 表示容忍程度。新的最值问题可以表述为：

$$\frac{1}{2}\|\vec{w}\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right) \tag{2.34}$$

约束条件由原来的$a_i \geq 0$变更为$0 \leq a_i \leq C$，但求解方法同前面介绍的训练方法类似。

特别的，在具体的问题中，常常使用变换不同的C值，来进行参数的范围估计的方法。通过C值的调整，在训练集上进行不同的训练，随后检查能够使训练集（或者开发集）获得最大准确率的C值作为最终的参数。

此外，支持向量机的两个基本问题中，还有一个我们始终没有涉及到，就是特征空间变换的函数$\phi(\vec{x})$。

通过计算的过程可以看出（无论是训练还是测试），我们所需要的计算都只限于样本点之间的内积运算。因此只需要定义任意两个高维空间中的向量的内积，即$K(\vec{x}_i,\vec{x}_j)=\phi(\vec{x}_i)\cdot\phi(\vec{x}_j)$就可以了。我们把这种内积函数称之为核函数（kernel function）。

核函数的选择是多种多样的。近些年来，不断地有研究人员在提出新的核函数。目前常用的核函数列举如下：

线性函数：$K(\vec{x}_i,\vec{x}_j)=\vec{x}_i\cdot\vec{x}_j$；

多项式函数：$K(\vec{x}_i,\vec{x}_j)=(\gamma\vec{x}_i\cdot\vec{x}_j+\gamma)^d$，$d>0$；

径向基函数（Radial Basis Function，RBF）：$K(\vec{x}_i,\vec{x}_j)=\exp\left(-\gamma\|\vec{x}_i-\vec{x}_j\|^2\right)$，$\gamma>0$；

双曲正交函数：$K(\vec{x}_i,\vec{x}_j)=\tanh(\gamma\vec{x}_i\cdot\vec{x}_j+\gamma)$。

在本文的后续实验中，采用径向基函数作为支持向量机的核函数。

2.6 说话人识别系统的评价指标

由图 2.1 知，说话人确认系统最后是将得分同阈值进行比较来判决该测试语音是属于目标说话人。因此其性能的好坏，取决于阈值设置的合理性。通常，我们选取的阈值往往和表征系统性能的 FAR 和 FRR 有关，在实验环境中通常将 FAR 和 FRR 相等时所对应的值作为系统的阈值。阈值越低，系统的 FRR 越低，说明目标说话人越容易闯入系统；阈值越高，系统的 FRR 越高，相应的 FAR 就越低，系统的安全性就越高，越不容易被闯入。但是在实际应用中，往往要根据应用的环境的不同来区别对待 FAR 和 FRR，比如，在对安全性要求相对不是很高的环境中，我们希望 FRR 会相对 FAR 要低，这样能增加系统的亲和性，使用户不容易产生反感的情绪。FAR 和 FRR 的计算公式如下：

$$FRR=\frac{n_{miss}}{n_{target}}\times 100\% \tag{2.35}$$

$$FAR = \frac{n_{fa}}{n_{imposter}} \times 100\% \tag{2.36}$$

其中，n_{fa} 表示非目标说话人被错误检出的次数，$n_{imposter}$ 表示非目标说话人实验的次数；n_{miss} 表示目标说话人未被检出的次数，n_{target} 表示目标说话人的实验次数。

2.7 本章小结

本章主要介绍了文本无关的说话人识别的基本原理。包括对说话人识别系统框架，以及框架中各模块的处理方法和常用算法的介绍。详细地对本文后面研究中需要使用的梅尔频率倒谱系数、高斯混合模型、支持向量机的定义、计算流程进行了介绍。

第3章 语音数据库和基线系统设计

3.1 引言

在第 2 章中，对文本无关的说话人识别技术的基础知识进行了介绍，本章对实验用的基线系统进行了说明，包括基线系统的参数设置以及实验所用数据库的信息介绍。同时对采用不同高斯混合数所得的实验结果进行统计，得出最优结果的参数配置作为后续实验的基础。

3.2 实验数据库及参数设定

3.2.1 实验数据库

实验中所采用的数据库，是在实验室环境下录制的一批针对嵌入式背景研究的语音数据。其主要参数如下：

人数：56 人，男女均衡。

文本：每人对 10 个中文词（约 2 到 3 个字）进行发音，每个中文词说 5 遍。

采样：16KHz 采样率，每个采样点用 16 位(bit)表示。

环境：录音设备为话筒；所有语音均在干净环境下录制。

对每个说话人，随即抽取一半的语音文件拿来训练，剩下一半的语音文件用来测试。

3.2.2 语音前端信号处理参数设定

表 3.1 语音前端信号处理参数设定

预加重滤波器	$1-0.95z^{-1}$
Hamming Window	$0.54-0.46*\cos(2n\pi/N-1) \quad n=1,\cdots,N-1$
帧长	32ms
帧移	16ms
三角滤波器个数	20

3.2.3 实验评价指标

本文是针对说话人确认系统展开的研究工作，根据 2.6 节介绍，说话人识别系统的错误率有两种：错误拒绝率 FRR 以及错误接受率 FAR。本文采用等错误率(Equal Error Rate，EER)来作为系统的评价指标，所谓等错误率就当错误拒绝率与错误接受率相等时的值。EER 的值越小，说明系统性能越好。错误拒绝率与错误接受率的计算在 2.6 节中有详细的介绍。

3.3 基线系统

3.3.1 基线系统结构

在说话人识别任务中，最常用的是将 MFCC 特征及其一阶差分 MFCC 特征联合使用来表征说话人的特性。在文[28][29]实验证明该特征参数比其他参数具有更优的识别性能。因此本研究的基线系统通过选取对说话人的 MFCC 及其一阶差分系数进行建模。

选取高斯混合模型作为基线系统的建模方法出于以下考虑：①从声学观点出发，每一个在高斯混合模型中独立的高斯分布来模拟说话人语音空间中某一种分类下的一项声学类别。这些类别分别代表说话人不同的发音状态，如鼻音、摩擦音等都是一些可能的类别；②对任一多类别的样本而言，高斯混合模型都有极佳的能力去近似这些样本(语音识别中为特征向量)空间的概率分布。

图 3.1 所示为一个典型的基于 GMM 的说话人确认系统结构图：

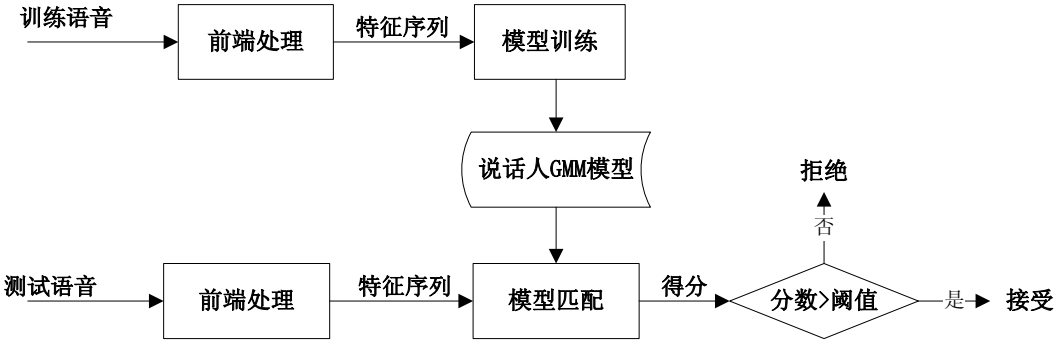

图 3.1 一个典型的基于 GMM 的说话人确认系统结构图

如图 3.1 所示，系统实现包含训练和识别两个阶段。

训练部分的具体步骤为：

①对训练语音进行采样量化。

②按照一定的帧长和帧移对训练语音信号进行加窗分帧操作。

③对训练语音进行前端处理。

④对每帧语音提取 MFCC 特征，组成 MFCC 特征序列。

⑤对所有 MFCC 参数用 GMM 建模，生成该说话人模型，存入模版库。

识别阶段的具体步骤为：

①对测试语音进行采样量化。

②按照一定的帧长和帧移对测试语音信号进行加窗分帧操作。

③对测试语音进行前端处理。

④对每帧语音提取 MFCC 特征，组成 MFCC 特征序列。

⑤将测试语音的 MFCC 参数同该测试语音所声称的说话人 GMM 模型进行比较，求出得分。

⑥将得分同经验阈值进行比较，得到最终的判决结果。

MFCC 和高斯混合模型的具体算法在第 2 章中有详细介绍，这里不再赘述。

3.3.2 高斯混合数选定

高斯分布的个数会对系统的性能有较大影响，若个数太少的话，则训练出来的高斯混合模型不能适当地描述说话人的特征，降低识别率；若个数太大的话，则可能出现稀疏预料的问题，还会大大的增加模型训练及辨认时的计算量。目前，对说话人识别系统还没有有效理论可以用来估算合适的高斯混合数。因此，为了选择最为适合的高斯混合个数，本实验分别选取了高斯混合数为 8、16、32、64、128、512 对说话人的 MFCC 特征进行建模，不同高斯混合数对系统性能的影响如下表 3.2 所示：

表 3.2 不同高斯混合数的系统 EER

混合数	M=8	M=16	M=32	M=64	M=128	M=512
EER	17.66%	12%	4%	2.67%	2.67%	3.7%

根据实验的结果，我们发现，高斯混合个数越大，说话人识别系统的整体表现越好。但是当混合数达到 512 时，系统的性能下降了。其可能原因是：32 维说话人特征空间向量用到 128 个时即达到饱和，稀疏语料引起了系统性能的下降。实验结果还显示采用 64 个高斯混合和 128 个高斯混合，系统的性能一样，但是 64 个高斯混合模型的训练时间要明显的少于 128 混合数的训练时间，因此本研究的后续实验对 MFCC 参数的建模都是采用 64 个高斯混合数。

3.3.3 实验结果

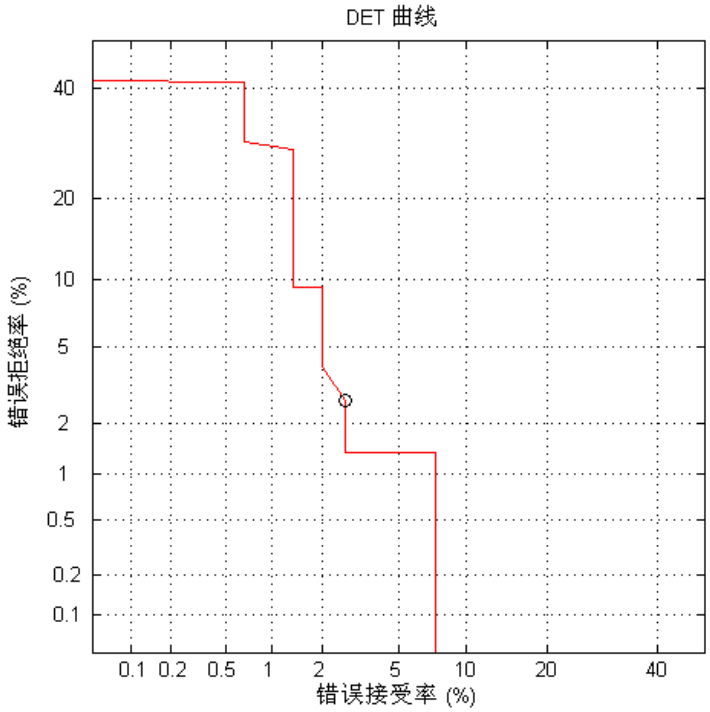

图 3.2 基线系统的 DET 曲线

图 3.2 中，横坐标代表错误接受率，纵坐标代表错误拒绝率。空心圆点表示错误拒绝率错误接受率相等时的位置，即 EER 的值。由图可以看出，基线系统的 EER=2.67%

3.4 本章小结

本章对实验所用数据库、基线系统结构进行了介绍。通过在相同的数据集上选用不同的高斯混合数分别进行实验，最后实验表明，当高斯混合个数为 64 时，系统性能最优。

第4章 文本无关的说话人识别快速算法研究

4.1 引言

将文本无关的说话人识别系统应用于嵌入式设备上有两个关键性的指标：①运行效率。对于用户的指令，系统能快速乃至实时地反应才能使用户获得最佳体验；②识别性能。足够高的识别性能才是其得以实用化的基础。本文将首先从提升系统的运行效率方面开展工作，进而提升运行效率的同时提升识别性能，其次从提升系统的识别性能方面开展工作，进而提升系统的性能和鲁棒性。

本章将对提升系统运行效率方面的工作进行详细的介绍，通过引入非线性分段的方法对第 3 章介绍的基线系统进行改进以减少系统时间开销，使其比基线系统更好的满足嵌入式应用的需求。

4.2 基于非线性分段的文本无关说话人识别

本论文的研究工作是在嵌入式背景下对文本无关的说话人识别系统进行研究，针对嵌入式平台相对运算速率低、存储空间有限的特点。我们引入了语音识别领域中常用的非线性分段（Non-Linear Partition, NLP）技术来改进基线系统。NLP 是一种将语音按照非线性方法划分为段的一种技术，通过压缩数据量来提高数据的处理效率[40]。

4.2.1 NLP 的思想和概念

非线性分段[41]，顾名思义，是将语音依据非线性的方法划分为段的技术。语音信号虽然是一个时变的随机信号，但是我们往往认为语音信号具有短时平稳性，即同一音素内的语音信号变化比较平稳，而音素间的变化则比较大。NLP 技术则是根据语音信号特征能量的变化情况，将特征序列划分为相对平稳的几段。它的具体描述如下：

假设一段语音有 T 帧，每一帧所对应的特征为 $x_i\left(1\le i\le T\right)$，则对特征序列 $X=\left(x_1,x_2,...x_T\right)$，定义 x_t 和 x_{t+1} 之间的距离为：

$$d_t=d_{cep}(x_t,x_{t+1})=\sum_{k=1}^{K}[W_k(x_t^k-x_{t+1}^k)]^2 \tag{4.1}$$

W_k 表示特征矢量第 k 维的权重，本实验所采用的$W_k=1$。假定将语音分为 N

段，那么平均每段的变化值为：

$$\Delta D = \frac{1}{N}\sum_{t=1}^{T-1} d_t \tag{4.2}$$

分段规则定义如下：

$$\sum_{t=1}^{n_{i-1}} d_t \le i * \Delta D \le \sum_{t=1}^{n_i} d_t \tag{4.3}$$

式子（4.3）中，$n_i\left(1 \le n_i \le N, n_0 = 0\right)$表示段分界点，对应于特征序列上的某一帧帧号。特征序列$(x_{n_{i-1}} + 1 \sim x_{n_i})$被归为第 i 段。

因此，我们可以看出，采用 NLP 进行分段后的语音信号，第 i 段的特征总变化量为特征总变化量的 1/N，即每一段的特征总变化量大致相等。这种算法的特点就是使得段内距离相对较小，而段间距离相对较大。

4.2.2 NLP 算法存在的问题

由前面介绍的 NLP 算法可以知道，NLP 算法其实是一种将以帧为单位的处理转换为以段处理的方式而节省系统处理开销的算法，因此以段为单位进行训练识别不如以帧为单位处理精细。为了保证说话人识别系统的精度，分段的合理性极为关键。然而在实验[42]中得出，该算法极容易受到微小干扰（如微小的辅音发音、杂音、尾音等）的影响，分段稳定性较差。

下图 4.1、图 4.2 分别是同一个说话人对同一个中文词的两次发音：

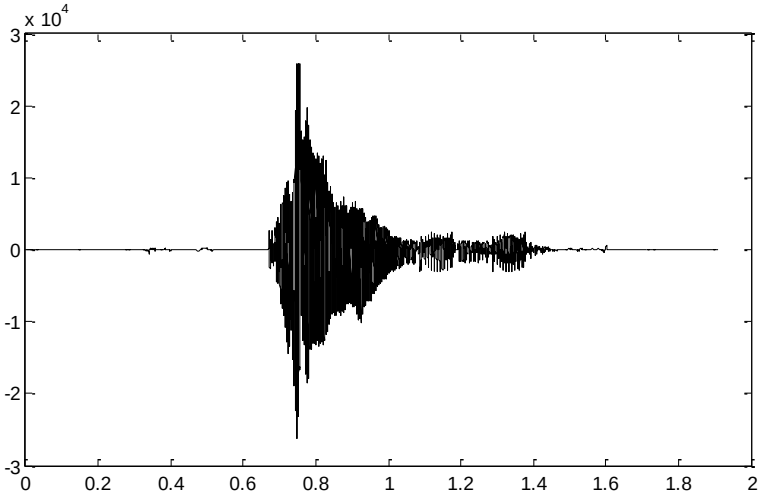

图 4.1　语音(a){n_1 = 34 n_2 = 6 n_3 = 1, n_4 = 70}

在图4.1中，将语音(a)划分为4段，分段的结果为{n_1 = 34 n_2 = 6 n_3 = 1, n_4 = 70}，其中$n_i\left(1 \le i \le 4\right)$分别表示第 i 段内包含的帧的数目。

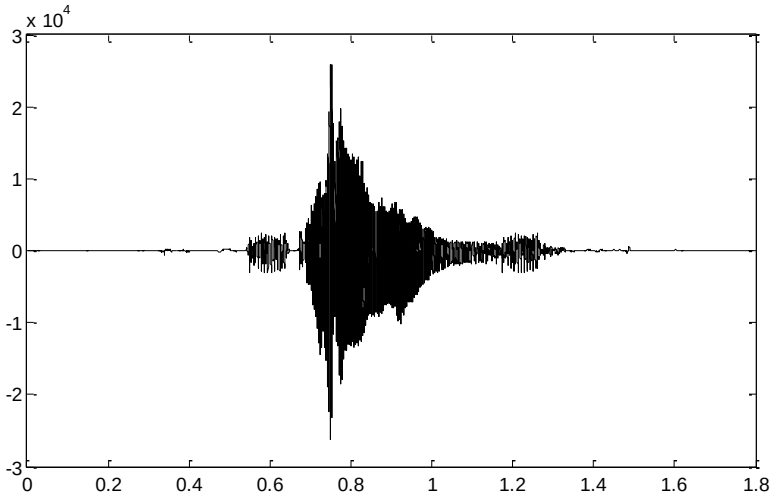

图 4.2 语音(b){n_1 = 29, n_2 = 14, n_3 = 36, n_4 = 32}

在图 4.2 中，将语音(b)划分为 4 段，分段的结果为{n_1 = 29, n_2 = 14, n_3 = 36, n_4 = 32}，其中 $n_i\left(1\le i\le 4\right)$ 分别表示第 i 段内包含的帧的数目。

图 4.1，图 4.2 中两个语音波形基本一样。只是语音(b)在最开始多了一小段杂音（辅音起始或微小噪音），且语音(a)和语音(b)均为 111 帧 $\left(n_1+n_2+n_3+n_4\right)$，但是这样分段的结果差距很大，显然传统的分段规则存在一些问题。

导致这种现象的原因在于，NLP 是基于距离累积的分段规则，受杂音帧的影响很大。杂音帧自身之间，以及杂音帧与其他帧之间的距离都较大，这些距离值将在计算中累积，从而影响到段的划分；而且这个影响还会传递到接下来的所有段的划分。在图 4.2 中的例子中，由于语音(b)一开始就受到杂音的影响，故所有段的划分都受到了影响。

4.2.3 改进的 NLP 算法

在引入 NLP 算法的系统中，我们对语音信号以段的单位进行训练和识别。这一过程相当于将一个大的问题进行了分解求值，它能有效地压缩数据信息，简化运算，提升数据处理效率。

根据前一节的分析，我们发现 NLP 受到杂音的干扰影响较大，算法鲁棒性较差。对此，我们提出了一些改进措施，NLP 算法中，对帧间的距离度量采用的是平方和距离，这种度量方式容易扩大干扰帧的影响力。根据文献[43]所述不加权的绝对值度量要优于不加权的平方和距离，因此本文提出采用绝对值距离度量来代替平方和距离度量。

即假设一段语音有 T 帧，每一帧所对应的特征为 $x_i\ (1\le i\le T)$，则对特征序列

$X=(x_1+x_2,...,x_T)$，定义 x_i 和 $x_j(1\le j\le T)$之间的距离为：

$$d_i=d_{cep}(x_i,x_j)=\sum_{k=1}^{K}\left|W_k(x_i^k-x_j^k)\right| \tag{4.4}$$

另外，考虑合理的分段应具有段内距离相对较小，段间距离相对较大的特点。假定某帧为段与段的分界点，则该帧前后帧的距离一定较大。引入清华大学语音语言实验室对分段规则的改进措施[42]，即假定某帧 *i* 作为分界点的分数 s_i：

$$s_i=\begin{cases}\dfrac{\left(s_i^{'}-\overline{s^{'}}\right)^2}{\sigma}, & if\left(s_i^{'}-\overline{s^{'}}\right)>0\\ 0, & else\end{cases} \tag{4.5}$$

其中 $s_i^{'}=\dfrac{1}{M}\sum_{n=1}^{M}d_{i-n,i+n}$

其中 $d_{i,j}$ 表示帧 *i* 与帧 *j* 之间的距离，*M* 为以帧 *i* 为中心选取的窗的宽度，$\overline{s^{'}}$ 和 σ则分别代表所有 s'_i 的均值和方差。一般来说，s_i 越大，表明第 i 帧前后帧的差异越大，则第 *i* 帧更有可能是一个段分界点。

由式(4.5)知道，该改进的方法以马氏距离为依据，取代了传统 NLP 中以距离累积为依据的分段规则。在该方法中，相邻分界点之间的所有帧被归为一段，而无需考虑这一段中具体累积了多少距离。这种分段规则对小的干扰是鲁棒的，而且也克服了距离累积规则所具有的传递性。

还是图 4.1 和图 4.2 中用到的语音(a)和语音(b)，选取分段数 *N*=4，采用改进的算法对语音（a）和语音（b）的分段结果分别是：{n_1 = 31 n_2 = 13 n_3 = 35, n_4 = 32} 和{n_1 = 20, n_2 = 18, n_3 = 34, n_4 = 29}。通过这两组数据对比，我们发现，采用改进后的 NLP 算法后，语音(b)同语音(a)的分段结果大致一样，符合我们预期的希望。另外，从分段结果可以发现，采用传统的 NLP 算法，语音(a)的第三段只有 1 帧数据，这样显然不合理，文献[42]分析造成这样的原因可能是语音短暂的停顿或者气音所引起的，而基于平方和的距离累积算法会对其进行覆盖。而采用绝对值距离累积显然更为可靠。同时，也说明分段的段数并不是段数越多系统性能越好，要根据实际的情况进行划分。

4.2.4 基于 NLP 的文本无关说话人识别系统

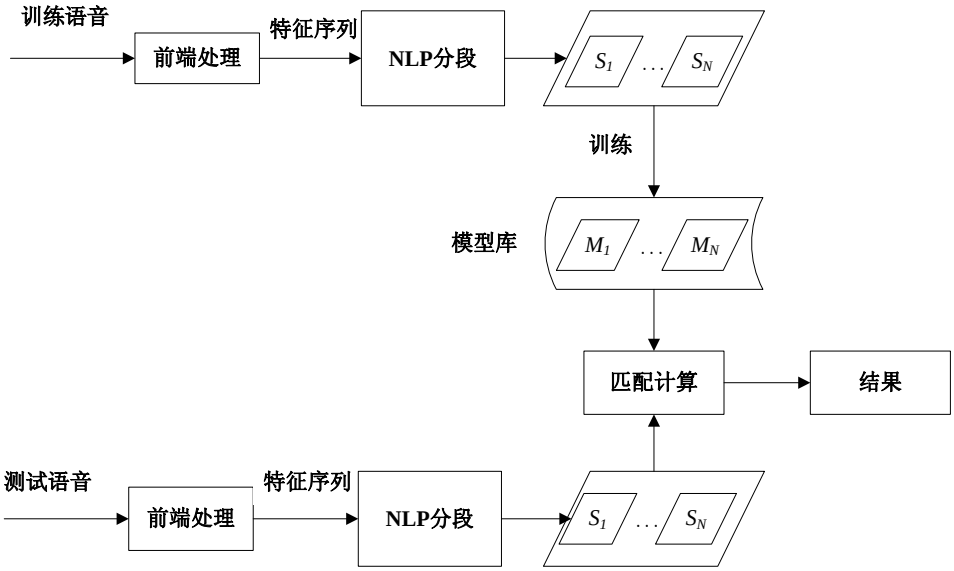

图 4.3 采用 NLP 的系统流程图

引入了 NLP 改进后的系统大致流程跟基线系统一样，只是在对特征参数进行建模前，先将特征序列按照 NLP 算法划分为几段，然后分别对每段进行建模，最后将这些基于段的高斯混合模型存入模型库，共同构成目标说话人的模型。基于 NLP 的文本无关的说话人识别系统也分为训练和识别两个部分。系统的流程图如上图 4.3 所示。

图 4.3 中，假定语音被分为 N 段，则分段后的特征序列为$(S_1,S_2,...,S_N)$,将目标说话人的训练语音的所有第S_N段特征采用 GMM 建模为 M_N，则目标说话人的模型为$(M_1,M_2,...,M_N)$。具体的步骤如下：

其中，训练阶段的具体步骤如下：

①对训练语音进行采样量化。

②按照一定的帧长和帧移对训练语音信号进行加窗分帧操作。

③对训练语音进行前端处理。

④对每帧语音提取 MFCC 特征，组成 MFCC 特征序列。

⑤根据 NLP 算法，将语音信号分为设定的 N 段。

⑥将目标说话人的所有训练语音的第 N 段 MFCC 特征序列进行建模得到目标说话人第 N 个模型，总共的 N 个模型构成说话人模型 M 。

另外，识别阶段的具体步骤如下：

①对测试语音进行采样量化。

②按照一定的帧长和帧移对测试语音信号进行加窗分帧操作。

③对测试语音进行前端处理。

④对每帧语音提取 MFCC 特征，组成 MFCC 特征序列。

⑤根据 NLP 算法，将语音信号分为设定的 N 段。

⑥将测试语音的第 N 段分别在对应的所声称的说话人的第 N 段的模型上进行打分，将最后得分进行归一化处理后即为该测试语音在所声称的说话人模型上的匹配分数。

⑦将最后的得分同经验阈值进行比较，得到最终的判决结果。

4.3 仿真实验和分析

4.3.1 分段数及高斯混合数的选定

表 4.1 不同分段数不同高斯混合数 EER

分段数	高斯混合数			
	M=8	M=16	M=32	M=64
1	17.66%	12%	4 %	2.67%
2	15.13%	10.61%	3.17%	3.84%
3	13.07%	7.47%	7.45%	8.43%
4	6.85%	6.85%	7.47%	9.47%

由于所选用的数据库语音长度只有 2~3 秒，过多的分段会导致一些段内语音帧数过少，反而会影响系统的性能。为了选择合适的分段数以及合适的高斯混合数，本实验采用没有改进的非线性分段算法对语音信号进行分段。由第 3 章的实验可知，针对实验中所用的数据库，当分段数为 1 的时候，系统采用 64 个高斯混合数能达到最优的性能，因此当分段数大于 1 的时候，系统要达到最优性能，对每一段分别建模的高斯混合数理应也不会多于 64，因此本实验选取了高斯混合数为 8、16、32、64，分别对分段数为 1、2、3、4 四种情况进行建模。引入了非线

性分段说话人识别系统的具体设计和详细步骤见小节 4.2.4。实验结果见上表 4.1 所示。

为了能直观的看清楚分段数以及高斯混合数对系统性能的影响，将上面的数据用直方图表示如下：

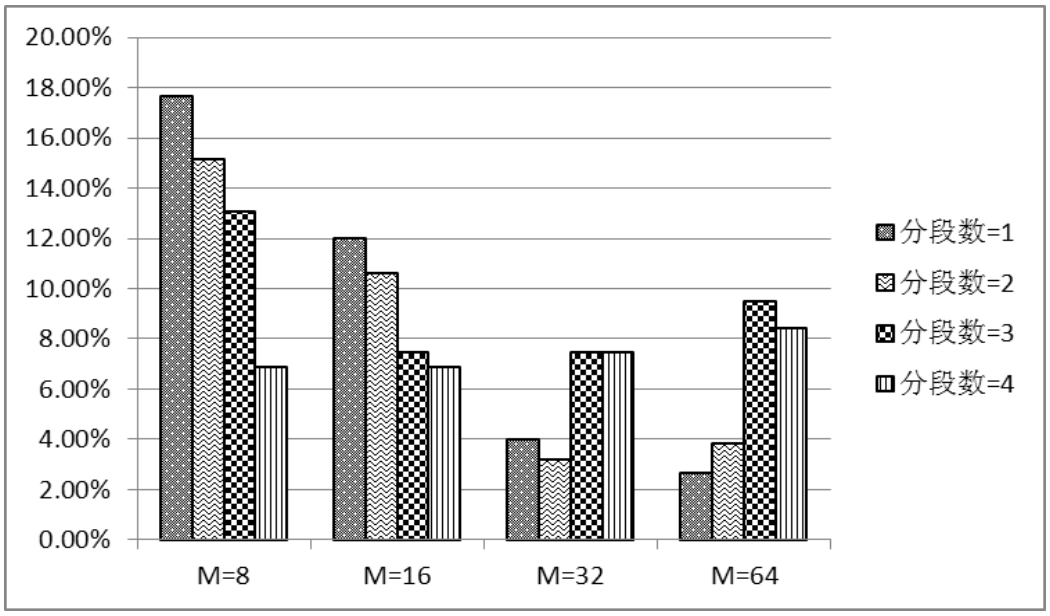

图 4.4 采用不同分段数以及不同高斯混合数的系统 EER 直方图

由图 4.4 可以看出，当高斯混合数为 32 的时候，系统的整体性能要优于其他混合数时的性能。但是我们也发现，在混合数为 64，分段数为 1（即基线系统）的情况下，系统的 EER 值最小，其次是混合数为 32，分段数为 2 的情况。引入了非线性分段的说话人识别系统在性能上并没有优于基线系统，造成这样的原因，在 4.2.2 节中已经对其进行了解释，是因为非线性分段算法极容易受到微小杂音的干扰，分段稳定性差。

因此，在后续的实验里，我们选取分段数为 2，每段进行建模的高斯混合数为 32，对改进的非线性分段算法对系统性能的影响进行验证。

4.3.2 实验结果

4.3.1 节中实验表明，采用非线性分段方法来对系统进行加速时，语音信号分段的段数设置为 2，高斯混合的个数设置为 32 是比较合适的。

(1) 非线性分段说话人识别系统

非线性分段说话识别系统的 DET 曲线如图 5.3 所示，由图可以看出该系统的

EER=3.17%，相对基线系统，系统性能下降了 19.08%。

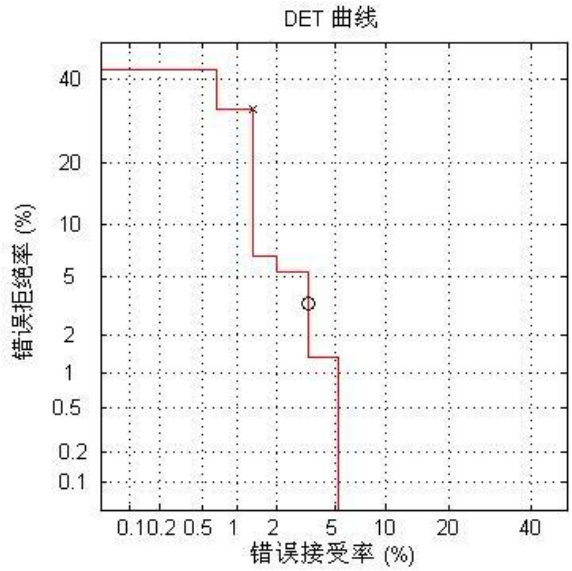

图 4.5 非线性分段的说话人识别系统 DET 曲线

(2) 改进的非线性分段的说话人识别系统

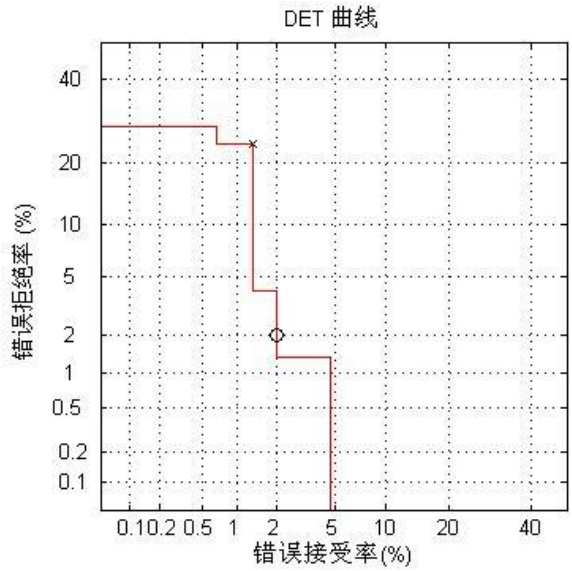

图 4.6 改进的非线性分段的说话人识别系统 DET 曲线

改进的非线性分段算法在 4.2.3 节中有详细的介绍，采用改进后的非线性分段的说话人识别系统的 DET 曲线如上图 4.6 所示，其 EER=2.13%，相对基线系统，系统性能提升了 20.22%

将以上结果统计如下表 4.2 所示：

表 4.2 不同系统的性能比较

系统	基线系统	基于 NLP 系统	基于改进的 NLP 系统
EER（%）	2.67	3.17	2.13
性能提升（相对值，%）	NULL	-19.08	+20.22

从表 4.2 中我们可以看出，虽然基于的 NLP 的说话人识别系统相对于基线系统的性能有所下降，但是改进后的 NLP 系统相对基线系统整体性能有所提升。另外，对分段后的语音进行建模时，高斯混合的个数大大的减少，也有效的减少了系统的训练识别时间。

4.3.3 实验分析

为了提升运算速度，本研究借鉴了非线性分段思想，使得语音信号变成在段的层面进行训练和识别，在这个方法里，每一个段均为多个帧的集合，通过这种方式可以压缩信息量，从而达到减少时间开销的目的。然而由于 NLP 算法是基于帧间距离累积的一种算法，它的分段规则抗干扰能力很差，一段微小的杂音即能导致差异巨大的分段结果。而分段结果的合理性和稳定性又直接决定了模型建立的精确性和识别匹配的准确性。因此，直接引入 NLP 方法，虽然时间开销降低了，但也带来了识别性能上的不小的损失。

本文提出的将帧间距离累积采用的平方和度量改为绝对值度量能够提升系统的一定性能，并引入采用基于马氏距离的分段规则，这种规则对语音中的微小干扰具有良好的鲁棒性并且保持了分段规则压缩信息量减少时间开销的优点。

4.4 本章小结

本章首先介绍了基于 GMM 文本无关的说话人确认系统的基本框架和系统流程，然后针对嵌入式平台的特点提出了采用 NLP 算法来压缩数据信息提高处理速度的方法，NLP 算法存在的问题进行了分析实验，最后介绍了分段规则的改进思路。通过实验验证改进后的 NLP 算法使得系统性能相对基线系统提升了 20.22%。

第5章 文本无关的说话人识别多特征融合技术研究

5.1 引言

噪音、信道差异和短语音等是说话人识别长期以来面临的几大问题。而嵌入式设备由于其计算能力、存储资源有限的特点而往往使用短语音对设备进行操作。另外，嵌入式设备便携性好的优点也决定了其所处环境的复杂多样性，噪声、信道差异问题严重影响着嵌入式说话人识别系统的性能。

一般来说，在研究说话人识别系统的鲁棒性时，我们无法兼顾所有的影响因子，很多情况下仅集中在消除某一方面带来的影响。近几年来，将韵律、词汇、音素等高层特征逐渐地应用于说话人识别系统中来代替直接消除某一种影响因子以增强系统性能成为时下一个研究热点。这样做的考虑基于以下几点：①这些高层特征参数在声学环境变化时能保持一定的稳定性，不易受信道噪声的影响；②这些高层特征还能反应说话人的说话习惯、风格等个性特征；③高层特征同基于底层的声学特征具有互补性，将高层特征同底层特征进行融合能增强系统性能。

本章将着重探讨如何将高层特征同底层特征进行融合的方法，其中包括提出了一种基于基频曲线的高层特征以及提出了采用SVM作为特征在分数域上的分类器融合工具的新方法。

5.2 分类器融合研究思路

分类器融合，就是对不同类型的多类特征，分别训练多个子分类器，随后将这些子分类器的决策结果结合在一起得出最终的决策结果。

从用于融合的单个分类器的输出来看，分类器融合方法可以分为三类：

①特征层，即在特征层进行混合。这种方法适用于所有子分类器的数学模型都相同，并且特征可以方便连接的情况。在语音识别系统等实际应用中，常见的方法是对每一帧语音提取多种特征，然后将它们直接连接起来，构成一个较长的特征向量，如将 39 维的 MFCC 特征和 12 维的 PLP 特征连接在一起构成一个 51 维的特征。但这种方法并不是对所有都适用，譬如语音帧数不等长时，没办法直接拼接，需要进行其他的处理。

②得分层，就是使用不同类型的特征，分别训练不同的子分类器，测试的时候让各个子分类器对输入的测试样本进行打分，给出“该样本属于每一个类别的程度”。可以给出对数似然分，或者是神经网络输出的概率等等，在最终判定的时

候对这些得分进行处理。

③决策层，即每个子分类器只给出一个辨识结果，然后在最终阶段采用投票等方法进行综合。

语音信号的不同特征参数代表的是人类不同的特性。理论上将不同的特征参数进行融合，能弥补各自的缺点达到提升系统性能的目的[44]。但是，如何将特征进行融合能既不掩盖自己本身的特性也能很好的同其他特征互相补充却是一个比较难的问题。目前常用的特征融合的方法就是直接特征拼接，但是这种方法只能针对特征维数相等的基于帧计算的特征，其应用比较局限性，效果也并不突出。

说话人识别系统里，不同类型的特征所反映的特性也是不同的。韵律特征具有不受信道影响，能反映说话人说话习惯和风格而越来越受到重视。寻找韵律信息最常用和最基本的方法就是求取语音的基频的特征，但是仅浊音帧上才有基频信息。在特征域进行融合，如果采用将基频信息直接拼接的办法会导致清音帧的数据丢失[45]，因此将韵律特征同导频谱特征直接拼接并不适合。在第 2 章中我们采用对基频曲线进行三次多项式拟合的方法获得基频曲线段特征，但是这种方法所求取出来的基频曲线段特征向量比较少，如果直接将基频曲线段特征直接同 MFCC 特征进行拼接的话，会造成 MFCC 对其的覆盖。所以选择在得分层对这两类特征进行分类器融合是一个较为理想的方法。

2.5.5 节中由对支持向量机的定义、训练、核函数的选择等具体介绍可知，SVM 本质上是一个线性分类器，由于引入了核函数的概念，使得支持向量机也能处理非线性判决的问题。它的中心思想就是调整判别函数使得它能最好的利用边界样本点的分类信息，能够很好的解决小样本情况下的机器学习问题。文[46][47]显示在口音辨识中引入支持向量机作为分类器融合工具能对系统的性能起到很大的提升效果。因此，本研究考虑到支持向量机的多种特性和优点，将其引入到说话人识别任务中，使用其作为我们的分类器融合工具。

5.3 多特征融合系统设计

5.3.1 系统整体框架

为了避免多特征混合时数据混叠，GMM 无法充分描述某类特征代表的说话人信息。我们分别对不同特征构建 GMM 子分类器，并采用不同类型特征分别在其对应的子分类器上进行打分，随后采用 SVM 对各子分类器上得到的分数进行综合而做出最终判决。我们设计的这种系统包括训练和识别两个过程。图 5.1 是我们提出的多特征融合系统的整体框图：

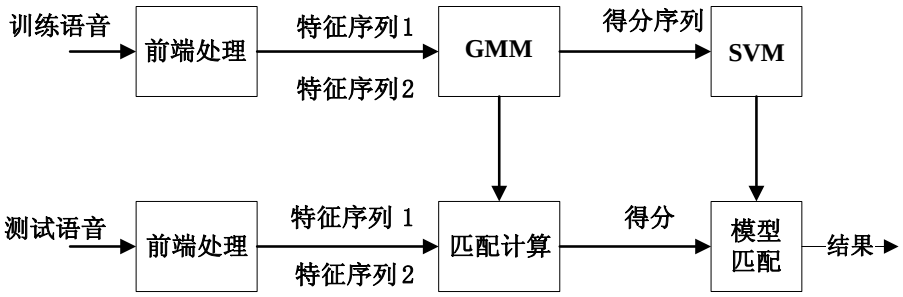

图 5.1　多重特征分数融合说话人确认系统整体框图

这样设计的好处在于：

①这两类特征的混合可以从不同的方面对说话人的特点进行覆盖。另外，这样的设计还具有扩展功能，可以根据实际需求选取更多的特征进行融合。多项式拟合方案的提出也能解决特征维数不等的问题。

②分别对每类特征进行建模，可以避免特征之间的相互混叠。采用 GMM 作为子分类器，将 GMM 的输出得分进行再利用，充分利用训练数据。

5.3.2　系统训练流程

该系统的训练过程，大概流程同传统的说话人识别系统流程一致。都是包括前端处理、模型训练、模型匹配、判决几个步骤。但是我们提出的这种多特征融合的方案在后期增加了一个 SVM 作为判决工具，同时，在各个子模块中也有细节性的改动。首先对目标说话人的所有 MFCC 特征，基频超音段特征分别进行 GMM 建模，然后使用训练集中的语音在这些 GMM 模型上进行打分，由于不同类型的特征得分范围相差甚远，为了避免得分范围大的覆盖范围小的，在分数层融合前还需要对分数进行归一化处理，本系统将得分归一到[-1,1]这个区间。图 5.2 是系统的训练流程图：

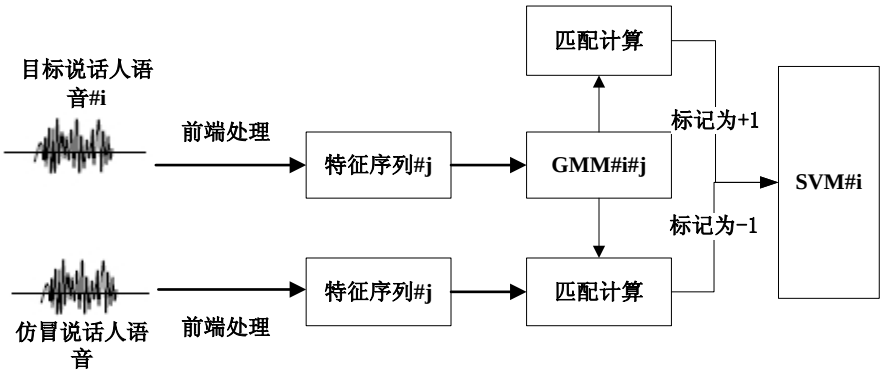

图 5.2　多重特征分数融合说话人确认系统训练过程

如图 5.2 所示，其中#i 表示目标说话人 i，#j 表示特征类型 j（本论文使用的特

征是 MFCC 和我们提出的采用多项式拟合方法拟合后的基频曲线特征）。我们将目标说话人#i 的特征类型#j 序列用 $\{Speak_i Fea_j\}$ 表示，进行模型训练得到 GMM#i#j，然后用 $\{Speak_i Fea_j\}$ 在对应的 GMM#i#j 上的得分进行归一化处理后标记为+1，将仿冒说话人在其对应的 GMM#i#j 上的得分归一化后标记为-1，将这两类得分作为输入参数用来训练目标说话人#i 的 SVM 模型。

另外，在具有 *N* 个说话人的说话人确认系统，对每个说话人而言，目标说话人只有一个，而仿冒说话人有 *N-1* 个。因此，在训练说话人的 SVM 模型时，也是一个二分类问题。我们对每个目标说话人训练一个 SVM 模型，即将目标说话人和 *N-1* 个仿冒说话人的得分一起训练。

假设该系统共有 *N* 个目标说话人和 *L* 种不同的特征，共训练 *NL* 个不同的高斯混合模型。对每一条训练集上的语音分别在这 *NL* 个高斯模型上进行似然分的计算，得到一个 *L* 行 *N* 列的矩阵。然后对矩阵的分数进行归一化处理，使之得分范围在[-1,1]之间。对上一步得出的 *L* 行 *N* 列的矩阵，取第 *N* 列作为第 *N* 个目标说话人的真实分数，其余的 *N*-1 列作为仿冒说话人得分训练说话人 *N* 的 SVM 模型。

训练的具体步骤如下：

①对训练语音进行采样量化。

②按照一定的帧长和帧移对训练语音信号进行加窗分帧操作。

③对训练语音进行前端处理。

④对每帧语音提取 MFCC 特征，组成 MFCC 特征序列。

⑤提取训练语音的 pitch 特征，对 pitch 曲线段进行三次多项式拟合得到维数为 4 的基频曲线段特征。

⑥对训练语音的 MFCC 特征构建 GMM 子分类器。

⑦对训练语音的基频曲线段特征构建 GMM 子分类器。

⑧将训练语音的 MFCC 特征在其对应的子分类器上进行打分，得到分数序列。

⑨将训练语音的基频曲线段特征在其对应的子分类器上进行打分，得到分数序列。

⑩将这两类分数分别进行归一化处理，是得它们的得分范围在[-1，1]之间。

⑪将归一后的分数联合成一个特征向量训练 SVM 模型。

5.3.3 系统识别流程

每次识别时，测试语音的特征数据也是先经目标说话人的 GMM 后，由其在 GMM 上的打分与目标说话人的 SVM 模型进行匹配计算，最终由 SVM 的输出来对该测试语音予以接受或拒绝的判决，如图 4.7 所示。

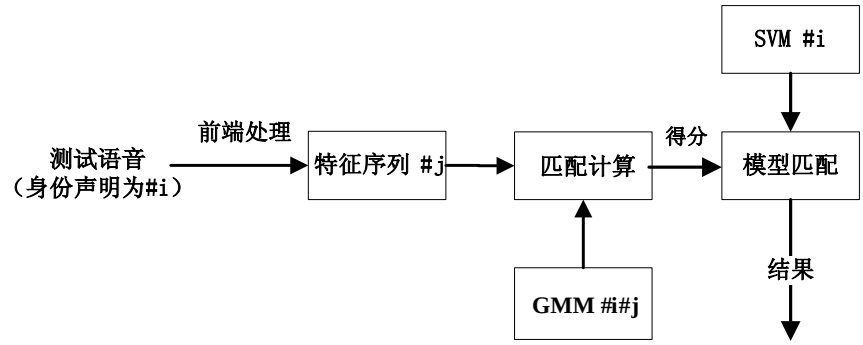

图 5.3　多重特征融合说话人确认系统测试过程

识别的具体步骤如下：

①对测试语音进行采样量化。

②按照一定的帧长和帧移对训练语音信号进行加窗分帧操作。

③对测试语音进行前端处理。

④对每帧语音提取 MFCC 特征，组成 MFCC 特征序列。

⑤提取测试语音的 pitch 特征，对 pitch 曲线段进行三次多项式拟合得到维数为 4 的基频曲线段特征。

⑥将测试语音的 MFCC 特征在其所声称的说话人对应的 GMM 子分类器上进行打分

⑦将测试语音的基频曲线段特征在其所声称的说话人对应的 GMM 子分类器上进行打分

⑧将两类分数进行归一化处理后连接一起。

⑨用训练过程中得到的支持向量机对上一步的向量进行测试，得到最终的判决结果。

5.4 仿真实验和分析

在第 4 章中，基于非线性分段的说话人识别系统性能已经得到有效的验证。我们发现在嵌入式平台上使用非线性分段算法是可行并且有效的。

针对嵌入式说话人识别系统性能对环境的依赖性过强的特点，其将不受信道环境因素影响能反应说话人韵律特点的高层特征与低层特征联合用于嵌入式说话人识别系统中以增强系统的抗噪性和鲁棒性。多特征融合利用特征间的互补性或者区分性能有效的提高系统的安全性能。但是多特征融合的系统往往具有数据存储量大、计算量显著增加的特点。所以多特征融合的系统在嵌入式设备上的应用非常有限，但是随着计算机技术的高速发展以及高性价比专用芯片的出现，在嵌

入式设备上实现多特征融合系统也将是必然趋势。本文对其的研究工作旨在希望对说话人识别技术乃至多生物特征融合技术方面的发展有一定的贡献。

下面将对本文提出的基于支持向量机的多特征融合的说话人识别系统的性能进行进一步验证。

5.4.1 不同种类单一特征的对比

基频曲线能反应说话人的声调特点，由于基频曲线不连续且不等长，为了能对基频曲线进行数学建模，本文采用了一种基于多项式拟合的基频曲线段特征。基频曲线段特征的具体的实现方法在第 2.4 节中有详细的介绍。由于本实验是对只使用基频曲线段特征的系统性能进行验证，因此在本实验中将基线系统的 MFCC 特征替换为基频曲线段特征，具体的实现步骤同基线系统一样。根据实验选择合适的高斯混合数，实验结果如表 5.1 所示：

表 5.1　不同高斯混合数系统 EER

混合数	M=4	M=6	M=8	M=16	M=32
EER	34.2%	28.1%	26.9%	26.9%	28.7%

由表 5.1 可以看出，当高斯混合数为 8 的时候，系统的性能最优，因此后续的实验都是采用混合数为 8 的高斯混合模型对基频曲线特征进行建模。

从特征方面而言，我们使用第 2.4 节中提到的 2 种特征，训练和测试流程依照第 2 章所述的高斯混合模型的训练和测试流程进行，由前面的实验将 MFCC 特征和基频曲线特征的高斯混合数分别确定为 64 和 8，实验结果如下表 5.2 所示：

表 5.2　仅使用单一特征，系统的 EER

特征类型	EER
基频曲线	26.9%
MFCC	2.67%

5.4.2 多特征融合的说话人识别系统

本实验，将之前得到的基频曲线段特征同基线系统采用的 MFCC 特征进行联合使用。融合的方案在 5.3 节中有详细的介绍，包括系统的整体设计框架，识别训练的具体步骤。

实验中的参数设置依旧采用前面实验所得最优解而选定的值。如，MFCC 特征采用 64 混合的高斯建模，基频曲线段特征则采用 8 混合的高斯建模。

下图 5.4 是多特征融合的说话人识别系统的 DET 曲线图，其中横坐标代表错误接受率，纵坐标代表错误拒绝率。空心圆点表示错误拒绝率错误接受率相等时的位置，即 EER 的值。由图可以看出，该系统的 EER=1.4%

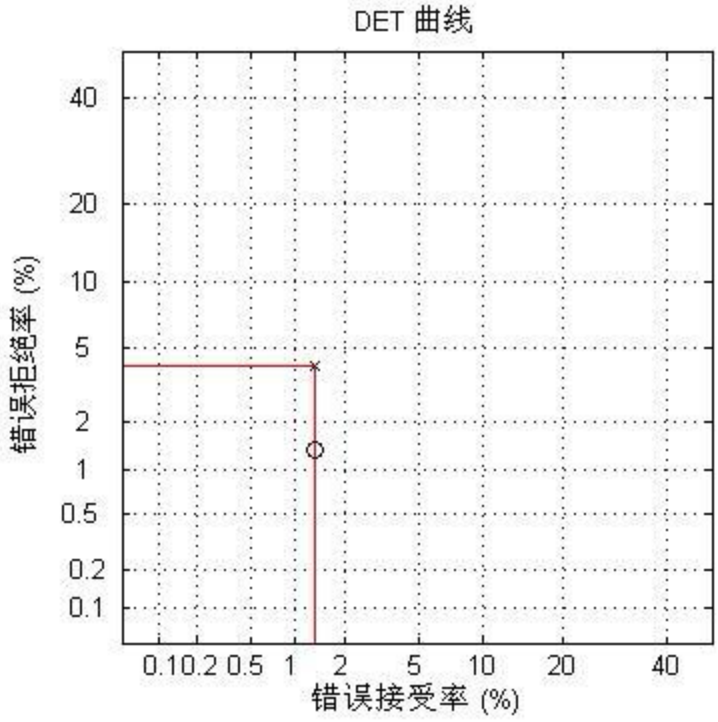

图 5.4　多特征融合说话人识别系统 DET 曲线

结合前面的实验，将所有结果统计到下表 5.3 中：

表 5.3　不同系统的性能比较

系统	基线系统	基于改进的 NLP 系统	多特征融合系统
EER（%）	2.67	2.13	1.4
性能提升（相对值，%）	NULL	+20.22	+47.57

由表 5.3 可以看出，我们提出的多特征融合系统性能最优，相对基线系统，性能提升了 47.57%。

5.4.3　实验分析

另外，表 5.3 表明采用提出的多特征融合后的系统性能最优。其主要原因除了

SVM 本身具有最优分界面的优点外，本文设计的方案是在子分类器上进行了第一次训练过程（高斯混合模型的训练）之后，在分数混合阶段又继续利用训练集的数据进行了第二次训练。因此，这种方法能够更加充分地利用训练数据，并且这种两次训练来分别调整参数和超参数的方法和思路，可以更有效地反映出具体任务和具体数据分布的特点，进而在测试集上也能得到一个相对较好的结论。

而使用单一的特征，低层特征的表现要好于高层特征。其原因有以下三方面：

①高层信息受到影响的因素较多.比如时变，以及人自身身体状态。

②单一高层信息不能充分表现说话人的特征。对一个语音文件来说，提取出来的高层特征的数量要远远少于语音帧的数量，不能充分描述说话人的特点。因此这些高层信息不适合单一使用，而适合作为说话人低层特征的补充。

③本实验所采用的数据库是针对嵌入式平台的特点而录制的，语音长度较短，所以我们提取出的基频曲线片段数量也较少，进而无法完全体现说话人的韵律特性，所以在单一使用基频曲线信息时实验的结果比较糟糕。

5.5 本章小结

本章介绍了在不同层次进行分类器融合的一些常见方法，并且对它们的优缺点进行了简要的说明。此外，针对说话人识别的特点，提出了采用支持向量机在分数层进行分类器融合的方案，并对我们所使用的系统框架进行了完整的描述。最后通过实验验证本文设计的多重特征融合方案则具有更优的性能，它相对于基线系统性能提高了 47.57%。

第6章 总结和展望

6.1 全文工作总结

随着语音识别技术在嵌入式市场的广泛应用，说话人识别技术由于其独特的优势受到了越来越多人的关注，将说话人识别技术拓展到嵌入式市场中去也具有了巨大的商业潜力。

本文就与文本无关的说话人识别系统研究做了两个方面的工作：一是引入了一种快速算法以提高系统的运行效率。二是采用多特征融合增强系统性能。研究的具体内容和成果如下：

①搭建了基于 GMM 的文本无关的说话人识别仿真系统。

②引入了非线性分段算法以提高系统的运行效率，并针对该算法的缺陷，对非线性分段算法进行了改进。

③采用了一种采用多项式拟合基频曲线段系数作为基频超音段特征的方案。汉语是一种声调语言，基频曲线变化能反应说话人发音的习惯，将基频曲线信息作为说话人特征补充到倒谱特征参数里面能提高系统的性能。但是由于基频曲线信息是不连续的且不等长的，为了能对其进行建模，我们引入三次多项式对其进行拟合，并用拟合的系数作为特征来描述基频曲线的变化情况。从实验的结果看来，单独使用这种基频段特征的效果并不太好，主要因为我们采用的语音数据库语音长度较短，这样导致能获取的基频特征数量较少，不能很好的覆盖说话人的个性特征。但是其作为 MFCC 特征的参数补充，确实能起系统性能增强的作用。

提出了一种基于支持向量机的多特征混合方案。该方法采用对多特征在其已有的模型上进行打分，将得分进行归一化处理后采用 SVM 在其分数域进行融合和判决。

6.2 工作展望

本文针对嵌入式平台下的文本无关的说话人确认系统展开了一些研究工作，提出了系统的加速算法，以及对如何增强系统的鲁棒性做了初步的探索。然而还存在着很多不足的方面。

①嵌入式设备环境多样性没有考虑进去。本论文所采用的语音数据库都是在干净的实验室环境下录制的。下一步的工作可以对数据库进行扩展，增加噪音环境下的语音数据。

②本论文所采用的语音数据库都是基于嵌入式平台的特点考虑的，所以数据库的语音长度都较短，能提取出来的基频信息有限。理论上是认为越长的语音越能反应人说话的韵律特性，也就是说，实验采用长语音应该能得到更好的效果。但是由于实验资源的局限性，没有对长语音进行研究。接下来的工作，可以录制一批长语音，对该特征融合的方案进行验证与改进。

③本论文在对增强系统鲁棒性的探讨中，只考虑加入了基频信息。但是否添加基频信息对无本无关的说话人确认系统是最优的选择呢。在后续的工作中，我们可以考虑加入共振峰等其他高层特征。

将说话人识别技术应用到嵌入式设备上是当前说话人识别领域的一个难点，尽管有很多研究人员孜孜不倦地为人类能用语音这种最自然的交流方式同机器进行沟通而努力，但是就具体应用到实际能广泛得到用户满意的理想状态来说，这项课题依旧还需要很长时间的探索和改进。

致 谢

衷心感谢郑方老师和李银国老师几年来给我的悉心指导与关怀，这两位老师无论在学习上还是生活上都给予了我莫大的帮助，并且在和这两位老师的相处交流中，他们无论是在学术上还是品德上，都让我十分敬佩，他们的教诲也必让我受益终身。在此，谨向两位老师致以最诚挚的谢意。

此外，感谢清华大学语音与语言技术中心的邬晓钧老师给予我的指导和帮助，感谢清华实验室的其他所有同学给予我的帮助和鼓励，特别感谢杨阳蕊博士给予我的帮助和建议。感谢重庆邮电大学汽车电子实验中心的程安宇老师、徐洋老师对我学习生活的关心与照顾，特别感谢吴渝老师对我论文修改提出的宝贵意见，吴老师渊博的学识、严谨的治学态度也是我学习的榜样，并将积极影响我今后的学习和工作。

感谢计算机学院夏淑芳老师，王宇同学，蒲甫安同学、舒适等同学对我的包容与照顾。

感谢我的父母家人，你们对我无私的爱一直是我前进的动力。

最后，衷心感谢在百忙中评阅论文和参加答辩的各位专家、教授！

签名：

日期：

参考文献

[1] 蔡莲红, 黄智德, 蔡锐. 现代语音技术基础及应用 [M]. 北京：清华大学出版社, 2003.

[2] 李财莲, 赵小阳, 王丽娟, 岳振军. 说话人识别中关键技术的现状与展望 [J]. 军事通信技术, 2005,26(2):62-65.

[3] S.Pruzansky. Patertn-Matching Procedure of Automatic Talker Recognition [J]. J.Acoust.Soc. Am, 1963(35):354-358.

[4] B.S.Atal. Automatic Speaker Recognition based on Pitch contourns [J]. J.Acoust.Soc.Am, 1972,52:1687-1697.

[5] J.Makhoul. Linear prediction:A tutorial review [J]. Proceedings of IEEE, 1975,63(5):561-580.

[6] Furui, S. Cepstral Analysis Technique for Automatic Speaker Verification [J]. Proceedings of IEEE Transactions on Acoustics, Speech and Signal Processing. 1981,29:254-272.

[7] Yegnanarayana, B., Prasanna, S.R.M., Zachariah, J.M., and Gupta, C.S. Combining Evidence From Source, Suprasegmental and Spectral Features for a Fixed-Text Speaker Verification System [C]. Proceedings of IEEE Transaction on Speech and Audio Processing. 2005:575-582.

[8] Soong, F., Rosenberg, A., Rabiner, L., et al. A vector quantization approach to speaker recognition. Internat [C]. Proceedings of Acoustics, Speech, and Signal Processing. 1985:387-390.

[9] Naik, J., Netsch, L., and Doddington, G. Speaker verification over long distance telephone lines. Internat [C]. Proceedings of Acoustics, Speech, and Signal Processing (ICASSP 1989). Glasgow, 1989:524－527.

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

[11] Haykin, S. Neural networks: a comprehensive foundation [M]. Macmillan, New York, USA. 1995.

[12] Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo, P. Support vector machines for speaker and language recognition [J]. Comput. Speech Lang. 2006, 20 (2–3), 210–229.

[13] 赵力. 语音信号处理 [M]. 北京：机械工业出版社, 2003.

[14] 杨阳, 陈永明. 声纹识别技术及应用 [J]. 电声技术, 2007, 31(02):45-50.

[15] 张广兰. 声纹识别的关键技术及发展趋势 [J]. 中国新技术产品. 2009(8):10.

[16] 朱浩冰，郭东辉. 声纹识别系统原理及关键技术 [J]. 计算机安全，2007(9):14-17.

[17] Gish. H, Schmidt. M. Text-Independent Speaker Identification [J]. IEEE Signal Processing Magazine, 1994,11(4):18-32.

[18] 郭皓婷. 基于声纹识别技术的应用难点研究 [C]. 2009 通信理论与技术新发展, 第十四届全国青年通信学术会议论文集, 2009.

[19] 于哲舟等. 智能仪器嵌入式声纹识别技术方法 [J]. 仪器仪表学报. 2004,25(4):447-450.

[20] 李济川等. 嵌入式语音识别及声控小车设计 [J]. 电子质量 ELECTRONICS QUALITY. 2003(11):11-12.

[21] Moon, Y., Leung, C., Pun, K. Fixed-point GMM-based Speaker Verification over Mobile Embedded System [C]. Preceedings of ACM Workshop on Biometrics: Method and Applications, 2003:53-57.

[22] Shin, K., Poon, J., Li, K. A Fixed-point DSP Based Cantonese Recognition System [C]. Proceedings of IEEE International Symposium on Industrial Electrons, 1995:390-393.

[23] 杨行峻. 语音信号数字处理 [M]. 北京:电子工业出版社, 1995.

[24] L.R.Rabiner, R.W.Schafer, Digital Signal Processing. Prentice-Hall [M], Inc, 1978.

[25] Chang, J., Kim, N. and Mitra, S. Voice Activity Detection Based on Multiple Statistical Models [J]. IEEE Trans on Signal Processing. 2006, 54 (6):1965-1976.

[26] Junqua, J., Reaves, B. and Mark, B.. A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize [C]. Proceedings of Eurospeech. 1991:1371–1374.

[27] 王炳锡, 屈丹等. 实用语音识别基础 [M]. 北京：国防工业出版社, 2005.

[28] Thian, N., Sanderson, C., Bengio, S. Spectral subband centroids as complementary

features for speaker authentication [C]. Internat. Conf. on Biometric Authentication (ICBA 2004), Hong Kong, China, July 2004:631-639.

[29] Kinnunen, T., Zhang, B., Zhu, J., Wang, Y. Speaker verification with adaptive spectral subband centroids [C]. Internat. Conf. on Biometrics (ICB 2007), Seoul, Korea, 2007:58-66.

[30] Raymond W. M. Ng, Cheung-Chi Leung, Tan Lee, Bin Ma, Haizhou Li. Prosodic Attribute Model for Spoken Language Identification [C]. Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing. Dallas, Texas, USA, 2010:5022-5025.

[31] 李庆扬, 王能超, 易大义. 数值分析(第四版) [M]. 北京: 清华大学出版社, 2001.

[32] 石柱. 声纹识别应用与矢量量化算法研究 [J]. 电声技术. 2006(10):44-47.

[33] Douglas A. Reynolds, Richard C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models [J].IEEE Transaction on Speech Audio Processing, 1995, 3(1):72-83.

[34] Douglas A. Reynold. Speaker identification and verification using Gaussian mixture speaker models [J]. Speech Communication, 1995, 17:91-108.

[35] Marc A. Zissman. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech [J]. IEEE Transactions on Speech and Audio Processing, 1996, 4(1):31-44.

[36] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning [M]. New York: Springer, 2001.

[37] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm [J]. Journal of the Royal Statistical Society, 1977, Series B (Methodological), 39(1):1-38.

[38] Christopher M. Bishop. Pattern Recognition and Machine Learning [M]. New York: Springer, 2006.

[39] Nello Cristianini, John Shawe-Taylor. An Intrduction to Support Vector Machines and Other Kernel-based Learning Methods [M]. 李国正, 王猛, 曾华军译. 北京：电子工业出版社,2004.

[40] Huang, X.-D., Cai, L.-H., Fang, D.-T., Ci, B.-J., et al. A Large Vocabulary Chinese Speech Recognition System [C]. Internat. Conf. on Acoustics, Speech,

and Signal Processing (ICASSP 1987), 1987:1167-1170.

[41] Zheng, F. Studies on Speaker-Independent Continuous Digit Recognition Methods and Chinese Speech Corpus [D]. Tsinghua University Department of Computer Science. 1992:28-29.

[42] Canhua Luo, Xiaojun Wu, Thomas Fang Zheng, Linlin Wang: Segmentation-based Method for Text-Dependent Speaker Recognition in Embedded Applications [C]. Asia-Pacific Signal and Information Processing Association, APSIPA ASC 2010.2010.

[43] 黄昌宁, 夏莹. 语音信息处理专论 [M]. 清华大学出版社, 1996.

[44] Eddie Wong, Sridha Sridharan. Fusion of Output Scores on Language Identification System [C]. The RTO IST Workshop on Multilingual Speech and Language Processing. Aalborg, Denmark, 2001.

[45] Bo Yin, Eliathamby Ambikairajah, Fang Chen. Combining Cepstral and Prosodic Features in Language Identification [C]. International Conference on Pattern Recognition. Hong Kong, 2006:254-257.

[46] Carol Pedersen, Joachim Diederich. Accent Classification Using Support Vector Machines [C]. IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007). Melbourne, Australia, 2007:444-449.

[47] Jue HOU, Yi LIU, Thomas Fang ZHENG, Jesper OLSEN, Jilei TIAN. Multi-layered Features with SVM for Chinese Accent Identification [C]. International Conference on Audio, Language and Image Processing (ICALIP 2010). Shanghai, China, 2010:25-30.