分类号 TP274.2 学校代码 10495
UDC 621.3 学 号 0915123025

武汉纺织大学

硕 士 学 位 论 文

复杂环境下特定说话人的语音识别研究

作者姓名: 张 琪

指导教师: 程建政 教授

学科门类: 工 学

专 业: 物理电子学

研究方向: 计算机语音信号处理

完成日期: 二零一二年三月

Wuhan Textile *University*

M. E. Dissertation

Research on Target Speaker Identification System under Noise Environment

By

Zhang Qi

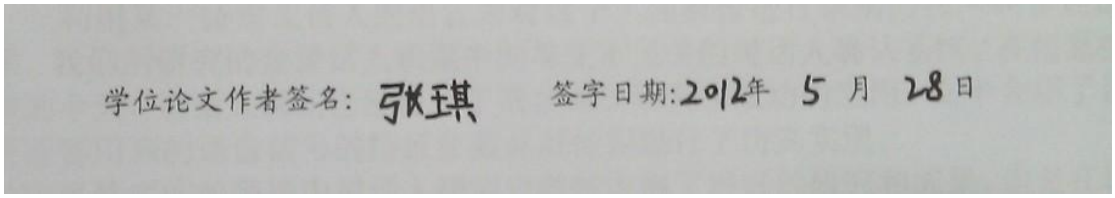
Directed by

Professor Cheng Jian-zheng

March 2012

独创性声明

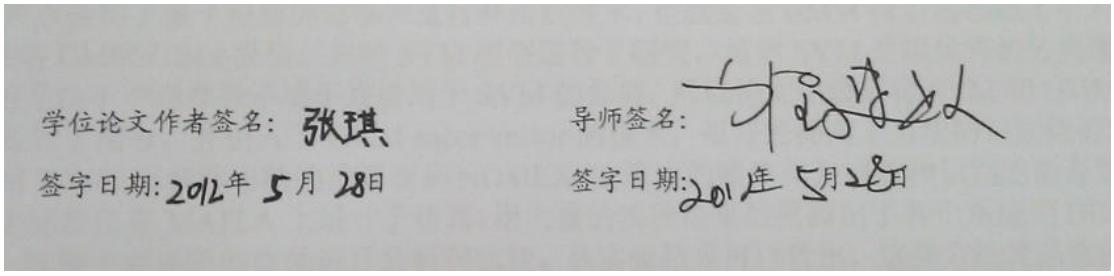
本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。



学位论文版权使用授权书

本学位论文作者完全了解 **武汉纺织大学** 有关保留、使用学位论文的规定。特授权 **武汉纺织大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）



论文题目：复杂环境下特定说话人的语音识别研究

专业：物理电子学

硕士生：张琪

指导教师：程建政

摘要

利用某一特定说话人的语音来对这个人的身份进行识别的技术叫做说话人识别技术。我们所研究的是说话人识别中的与文本无关的说话人确认系统。在信息技术高速发展的今天，说话人确认已经得到了很全面的研究和成功的应用。文中介绍了说话人识别所需要用到的语音信号的特征参数并对他们进行了仿真实现。

虽然在实验环境中说话人研究已经能取得了很好的研究和成果，但是在运用到现实生活中的时候，因为复杂的环境噪声的影响会造成系统识别性能的急剧下降。本文针对这个问题，分为两个方向研究了如何提高说话人识别系统在环境噪声下的鲁棒性的方法：

首先介绍了基于特征参数的噪声鲁棒性算法，在这类算法里面着重研究分析了 Delta 参数、谱减法、PCA 和 RASTA 滤波等在说话人识别中常用的消除噪声影响的技术。用 HTK 工具箱分别对语音信号提取 MFCC 参数和它的 Delta 参数、Delta-Delta 参数，用来对语音信号进行识别，发现 Delta 参数可以提高系统的识别性能。在介绍谱减法时使用谱减法对语音进行了增强，由于谱减法对语音加强之后会出现“音乐噪声”，引入了改进的谱减算法。PCA 可以对参数进行降维和去除一部分噪声，本文对 PCA 主轴和数据方向的一致性、PCA 变换对数据的扩展性能和数据 PCA 的降维有益于 GMM 模型分类等方面进行了验证。在对 RASTA 的研究过程中，将其用于 PLP 参数的滤波，在实验中，对干净和带噪语音信号分别进行 PLP 和 PLP-RASTA 特征参数的提取，发现干净语音和带噪语音所产生 PLP-RASTA 参数的谱图要比两者的 PLP 参数谱图相近，证实了 PLP-RASTA 参数的鲁棒性。并基于特征参数的融合提出了一个新的抗噪特征参数，并用实验数据证实了它的有效性。

然后介绍了基于模型的对噪声进行补偿的技术，也就是在 GMM 模型的基础上引入 UBM 的 GMM-UBM 模型。并对 SVM 模型进行了研究，虽然 SVM 是很优秀的分类模型，但是由于声道参数不适于直接用于 SVM 的分类，所以论文中最后将 SVM 和 GMM 模型进行了混合，并引入了 GMM super vector 的技术，很好的提高了系统的识别性能。本文用 TIMIT 语音库中的纯净语音和 NOSIEX-92 库中的噪声作为训练和识别的语音数据对上述算法在 MATLA 上进行了仿真，用大量的实验结果数据画出了各个系统的 DET 曲线，以便于对系统的性能进行分析和比较。从实验结果可以看出，这些方法使系统的噪声鲁棒性得到了很大的提高。

关键词：说话人确认；GMM；GMM-UBM；SVM；MFCC

研究类型：应用研究

Subject：Research on Target Speaker Identification System under Noise Environment

Specialty：Physical Electronics

Name：Zhang Qi

Instructor：Cheng Jian-zheng

ABSTRACT

Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. In this article the work leading to this thesis has been focused on establishing a text-independent speaker verification system. Speaker verification is to verify a person's claimed identity.

In spite of the speaker verification has achieved satisfactory results in clean speech environment, those systems' performance decreases drastically under the channel or noise environment. In exploration of these issues, many different methods are implemented, along with two kinds of noise robust technologies: noise robust technologies based on characteristic parameters and based on models. Systems were built based on those two kinds of noise robust technologies for the purpose of improving the recognition accuracy.

Firstly, noise robust algorithms based on the characteristic parameters are introduced. In this article, studies and analysis of elimination noise impact based on characteristic parameters are mainly focuses on Delta, SS (Spectral Subtraction), PCA, RASTA filtering etc. MFCC and its Delta, Delta-Delta parameters of voice signal are extracted and used for the recognition in HTK Speech Recognition Toolkit, the test results shows that Delta parameters can improve the recognition performance. Noisy speech were enhanced using spectral subtraction after the introduction of spectral subtraction, due to speech enhancement based on spectral subtraction followed with the" music noise", improved spectral subtraction algorithm is introduced. PCA, which dimensionality reduction and denoising on characteristic parameters, in this article, the consistency of the PCA spindle and the direction of data, expansion properties of PCA transformation to the data and GMM model's benefits from data dimension reduction by PCA are verified. In the research on RASTA, it is mixed with PLP,the PLP and PLP-RASTA are extracted, by observe the spectrums, we can find that the PLP-RASTA feature parameters of clean and noisy speech are much closer than the PLP feature. A new anti-noise characteristic parameter is proposed based on the feature parameters fusion and the experimental data have confirmed its validity.

Then, in order to elimination noise impact, robustness techniques based on models are introduced. as a outstanding compensate model, GMM-UBM has been substantially used in text-independent speaker verification. And SVM is also introduced, SVM is a very good classification model, However, due to the unsuitable of channel parameter directly used into SVM classification, SVM and GMM models are mixed, and the introduction of on GMM

super vector technology can improve the performance of the system effectively. Results of verification tasks on the TIMIT and NOSIE-92 databases on MATLAB show that increased model fit directly with using those noise robust technologies. It is shown that both of these approaches improve the performance of system trained on speech data.

Keywords: speaker verification; GMM; GMM-UBM; SVM; MFCC

Thesis: Theory Research

目 录

1 绪论

1.1 背景

作为人类交流和交换信息的工具，语言无疑是其中最重要、最方便、最直接的工具。也是人类特有的特征，而语音是语言的表现形式。实现计算机和人类能直接通过语音进行交流是人类自计算机诞生以来不懈追求的一个梦想。随着信息技术的发展，语音信号处理得到了得到了广泛的应用和发展，应用方向也越来越多，语音信号处理的若干研究方向如图1.1所示[5]。

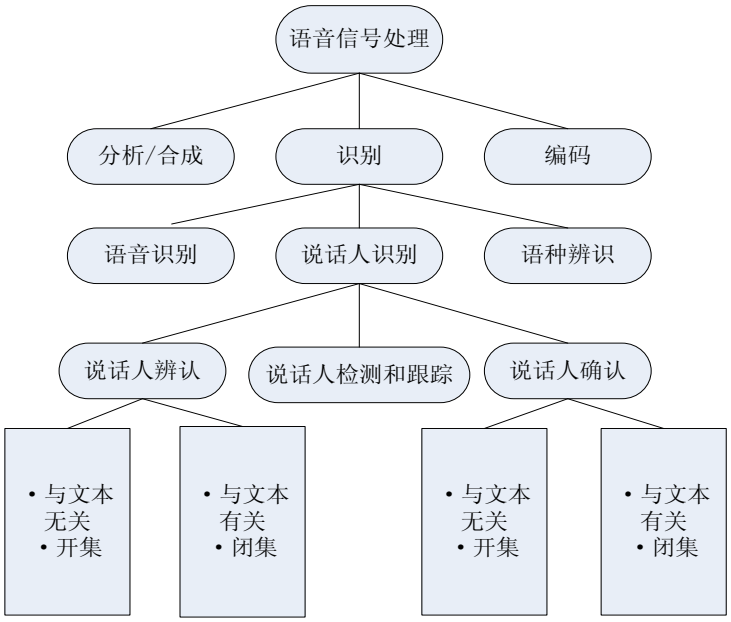


图1.1 语音信号处理

语音识别是指计算机对人类语言进行正确响应的技术[6]语音识别的目的是从语音信号中利用相应算法从语音信号中自动的提取出人们感兴趣的信息，语音识别有语义识别、语种辨识以及说话人识别等研究方向。语义识别就是一般所说的语音识别（Speech Recognition），就是识别出语音的内容，从而根据其信息，执行人的各种意图；语种辨识或称语言辨识（Language Identification）是通过分析处理一个语言片段以判别其所属语言的语种；说话人识别（Speaker Recognition）并不注重语音信号中的语义内容，而是对语音信号进行处理，根据语音里包含的说话人的个人特征信息从而对对说话人进行

身份鉴别与认证的一种技术。

说话人识别和语音识别一样，都是对采集到语音信号信号经行处理，提取语音信号的特征参数，建立相应的识别模型，以此为标准来进行识别。两者的区别在于语音识别是只关心语音中所包含的语义内容，力图消除不同说话人的区别，在不同人的语音信号之间寻找共通之处，而说话人识别却是需要从语音信号中提取出能代表说话人个性特征的参数，强调不同人之间的语音信号的差别，并不很关心语音包含的语义内容。

1.2 说话人识别

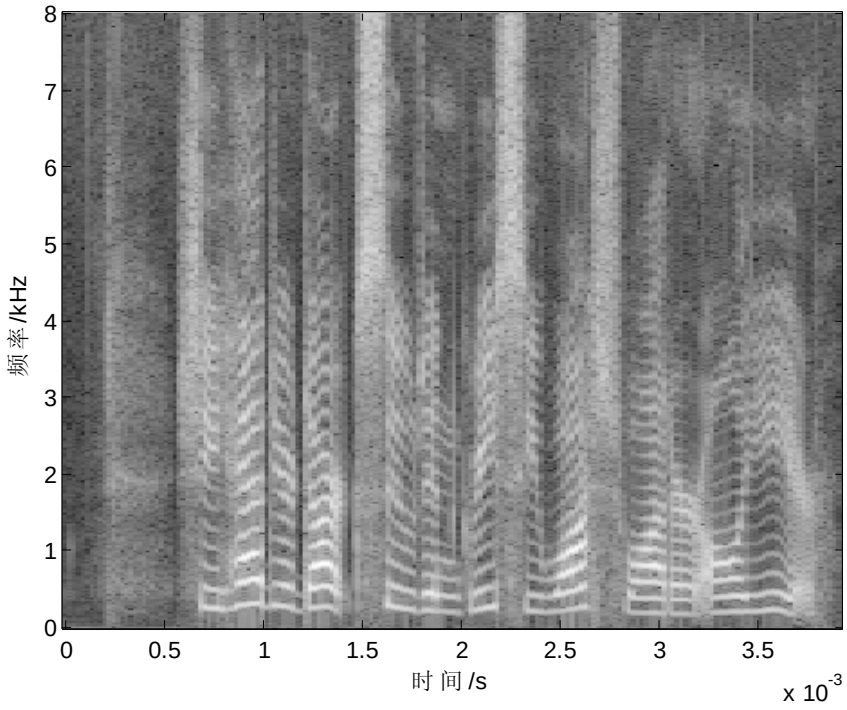


图1.2 一段语音的语谱图

当今社会信息技术的发展使得生物识别技术日渐成熟并在人们的生活中具有重要的作用。所谓生物识别技术，就是提取人体的固有生理特征和行为特征，利用计算机等高科技方法进行识别。现在主要的生物识别技术有：指纹识别、人脸识别、虹膜识别、还有本文要介绍的语音识别中的说话人识别，也叫做声纹（voiceprint）识别。1945年，Bell 实验室的Lawrence G. Kersta 通过研究用语图仪（Sonograph）绘出的语谱图（Sound Spectrograph），他们发现了同一个人发同一个音的谱图，总是比其他人发这个音得到的语谱图更加地接近。谱图直观明了，十分类似于指纹。由此提出了声纹的概念。生物识别技术比起传统的身份识别方法更加方便，它不容易丢失、遗忘或者被盗。目前这些技术广泛应用于许多领域中，如财经领域、安全保卫领域、公安司法领域、军事领域、信

息服务领域等。

语音是人的自然属性之一，由于说话人发音器官的生理差异和后天形成的行为不同，事实上，每个人的牙齿、舌头的尺寸，口腔和鼻腔的长度，声带的形状和韧性，声道的大小和形态等方面都存在着很大的差异。这些因素使得每个人的发音特征和习惯都有所不同，具体就反应在声音信号中。所以通过对语音信号进行分析处理，就能够找出其特征，分辨出说话人。说话人识别与指纹识别、虹膜识别等生物特征识别相比，它的优势在于输入、识别所用设备简单，一般只需麦克风和普通计算机等，识别也只需要说几句话，系统成本低且采集方便。另外利用电话网络实现远程客户服务。这项技术使得可以通过说话人的声音来对说话人的身份进行确认或辨认进行服务如语音拨号、电话银行、电话购物、数据库访问服务、语音邮件、机密信息的安全控制和远程访问计算机等成为可能。

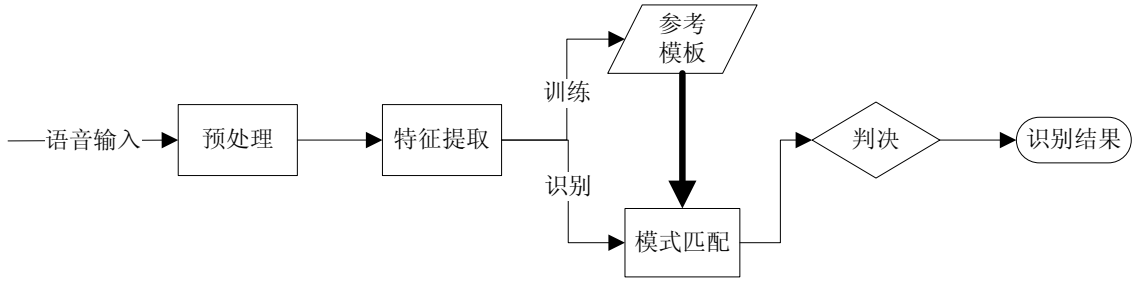


图1.3 说话人识别系统框图

一个说话人识别系统如图1.3所示，它的建立有两个重要的阶段：训练阶段和识别阶段。训练阶段需要使用者的若干训练语音片段。提取这些语音片段的特征参数以作为标准对系统进行训练学习，建立模板或模型参数参考集。在训练阶段时，截取待识别者的语音片段，对其进行特征参数的提取，然后参照建立的模板或模型参数参考集进行比较，根据一定的相似准则进行判定。

1.2.1 说话人识别的分类

按不同的角度，说话人识别有多种不同的分类方法。自动说话人识别（Speaker Recognition, ASR）是一种自动识别说话人的过程。按其最终目标可分为四类：

（1） 说话人确认（Speaker Verification, ASV）：判断一段未知语音是否来自于一个特定人的语音片段，只需要输出“是”或者“否”，是一个二元问题。

（2） 说话人辨认（Speaker Identification, ASI）：判断一段未知语音是来自于N个模型中的哪一个人所说的语音片段，多选一问题。

（3） 说话人检测（Speaker Detection, ASD）：指对一段包含多个人说话的语音，判断语音中是否含有特定说话人的语音。与说话人确认相似，也属于二元检测，只需要

输出是或者否。区别只在于给定的语音是属于多个人的。

（4） 说话人跟踪（Speaker Tracking, AST）：在说话人检测的基础上，如果该多人语音中包含了某个特定说话人的语音，则标出此特定说话人语音部分在多人语音中的位置。

也有文献上把（3）和（4）归为一类，称为说话人探测跟踪（Speaker segmentation and clustering）[3]。

当说到说话人识别时传统的主要的目标是指前面两种，即说话人确认和说话人辨认。两者的虽然在可选决策数目上不同，但说话人确认和说话人辨认，在本质上没有很大的区别，都是根据说话人所说的测试语句或者关键词语，从中提取说话人本人特征相关的信息，与训练好的参考模型比较之后做出判断。说话人辨识是需要将待识别的语音判别为已注册人之中的哪一个人所说的，这个是多选一的问题。如果已知待识别的说话人是在已注册的说话人集合内，这就是闭集测试（close-set），也就是识别系统已经先验得说话人属于某参考说话人集合；反之，如果不知道该说话人有没有注册，就是开集测试（open-set）。对于开集测试，增加一个确认过程，辨识系统需要首先对语音进行确认，判断该说话人是否已经注册，并据此接受或拒绝辨认结果。显然，多了这个确认过程就会导致整个说话人辨识的效率及性能下降，开集识别的难度要比闭级大很多。说话人确认是判别待识别的语音是否与他所申明的已注册模型相符，系统只做出“是”或“不是”的二元判断。两种系统的基本原理如图1.3所示。事实上，两种系统的主要区别是判断的数目不同，说话人辨认是在全部注册说话人的范围内进行的，需要进行多次判断，因此辨认性能与注册的人数有关。随着注册人数的增加，不仅训练时间变长，各个用户之间就变得更难得区分，辨认性能力会有所下降；而说话人确认与注册的人数关系较小，随着注册人数的增大，确认性能不随用户的增加而变化，接近稳定的常数。

在进行说话人识别时，根据限定的语音内容，还可以将其分为三类：与文本无关（Text-Independent）、与文本有关（Text-Dependent）和文本指定型（Text-prompt）。

（1） 文本相关是指系统训练时精确的建立了说话人的模型，识别时要求训练语音和测试语音所对应的文本一致，要求说话人按规定的内容发音。这种识别模式的建模方法比较简单，所需要的训练语音与测试语音的长度较短，但由于对文本内容的限制，通常只在出入境管理、安全认证等及身份认证相关方面使用。

（2）文本无关是指不测试结果与训练语音和测试语音所对应的文本是否一致文本无关，对测试的内容不做限定，说话人可以不用确定说话内容。模型的建立相对与文本相关的建立要困难一些，由于对文本内容限制较低，使用方法要更加灵活，用户使用方便，应用区域较多，特别适合在国防侦听、刑事侦查等与语音侦听相关方面使用。通常来讲，文本相关的说话人识别精度要比文本无关的说话人识别要高，但是文本无关使用的灵活性相对于前者要好得多。如果是从安全性上来说，显然只有前两种情况是不完全

的，如果事先把使用者语音录下来的话，就会有装置误接受的危险。所以有了下面的第三种类别。

（3） 指定文本是要求测试语音所对应的文本为系统所设定文本(不一定与训练语音的文本相同)的说话人识别。指定文本的说话人识别可以分成两种方式：一种是在进行识别时，系统将随机地指定要求说话人说出的文本，当说话人所说的文本与指定的文本相同，并且说话人所说的被认可时，系统才接受该说话人；另一种是在进行识别时，系统随机的对说话人提问，这些问题是预先设置的一个或几个问题，当说话人回答的文本与预先指定的答案的文本相同，同时说话人所说的被认可时，系统才接受该说话人。一般情况下，指定文本的说话人识别与语音识别技术结合使用。通常来说，与文本相关的说话人识别需要的训练语音时间较短，由于系统性能高从而很容易建立模型。但是识别与文本无关的说话人，需要很长时间的语音训练，并且在测试时，需要的语音时间较长。本文主要研究与文本无关的说话人确认。

1.2 .2 说话人识别的模型的分类

（1） 模式匹配法：模板匹配法的要点是，在训练过程中从说话人发出的训练语句中提取相应的特征矢量，这些特征矢量能够充分描写各个说话人的个性特征。这些特征矢量称为各个说话人的模板。在测试阶段，按同样的方法在说话人的测试语音里面提取测试模板，根据与相应的参考模板相比较得到匹配程度也就是模板之间的距离来做出判断。

① 动态时间规整方法（Dynamic Time Warping, DTW）。

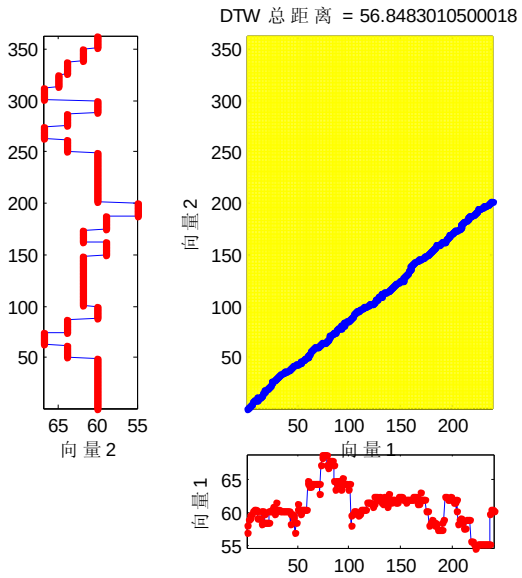


图1.4 动态时间规整算法

这是在80年代的与文本相关的语音识别和说话人识别中提出的一种算法，在固定短语的应用中效果较好，主要是利用动态时问规整以对准训练和测试特征序列来进行判别，这个方法的实现很简单，图1.4所示；由于直接在时间序列上进行，受噪声，信道的干扰很大，应用范围受到限制，在概率统计模型算法成熟之后，目前已经很少有研究机构采用这种方法了。

② 矢量量化(Vector Quantization, VQ)方法：矢量量化技术是最早是用于聚类分析的数据压缩编码技术。不直接在时域上进于预处理，对倒谱参数进行聚类，把每个人的特定文本训练成码本，识别的时候根据类别失真度进行判别，算法复杂度不高，且识别精度并不低。目前主要使用方法是作为其他方法方法的一种初值处理方法,

（2） 概率统计方法[17]：说话人的语音信息在较短时间内可看做是平稳信息，通过对稳态特性的统计分析，根据特征参数的概率分布建立模型，然后可利用均值、方差等统计量和概率密度函数进行分类判决，适合于文本无关的说话人识别。

① 隐马尔可夫模型(Hidden Markov Models, HMM)方法：是一种基于转移概率和输出概率所建立的随机模型，可以描述语音随时间变换的情况。有多种方法，最初的应用是从Forward、Backward算法的得分中进行判别，目前也有用HMM进行LVCSR后从音素层面上进行建模判别的，比如建立基于每个音素的GMM模型。HMM对噪声的鲁棒性较低，训练时计算量较大。

② 混合高斯模型(Gaussian Mixture Model, GMM)：GMM模型实际上是一种单状态的HMM，通过用多个高斯分布的线性组合来近似多维矢量的连续概率分布，有效地刻画了说话人的特征。其中GMM-UBM模型是目前主流算法，在与文本无关的说话人识别中效果比较好。

（3） 辨别分类器方法

① 人工神经网络方法（Artificial Neural Network, ANN）：神经网络在某种程度上模拟了生物的感知特性和信息处理机制，能把大量结果简单的计算单元互相连接，是一种分布式并行处理结构的网络模型，有很多种形式，如多层感知、径向基函数(RBF)等，具有自组织和自学习的能力、很强的复杂分类边界区分能力及对不完全信息的鲁棒性，其性能近似理想的分类器，可以显式训练以区分说话人和其背景说话人，但是在最初的说话人识别中应用，训练量很大，训练时间长，随着识别人数的增加训练难度也增高，模型的可推广性不是很好。

② 支持向量机（Support Vector Machine，SVM）的分类方法：主要有两类输入，一类是采用多项式方法对声学特征参数进行扩展，第二类是与GMM方法相结合，也就是本文会介绍到的GMM-SVM。SVM方法有较高的精度，型存储和运算过程中对于存储空间和内存的要求都很大，另外对于短时语音的情况存在着性能不稳定的缺陷。

（4） 混合方法：把以上分类方法与不同特征经行有机组合可显著提高说话人识别

的性能。本文会介绍到 GMM 和 SVM 的混合方法。

1.3 说话人识别的系统性能评价标准

1.3.1 ROC 曲线

假设s为未知的语音片段来自系统使用者的声音的状态，n为不是系统使用者的状态，识别结果只有s和n两种情况，在不同状态下获得的这两种概率结果有且只有s|s，s|n，n|s，n|n四种情况，将他们的分布概率分别记为P(s|s)，P(s|n)，P(n|s)，P(n|n)，有：

$$\begin{cases} P(s\,|\,s)+P(n\,|\,s)=1 \\ P(s\,|\,n)+P(n\,|\,n)=1 \end{cases} \tag{1.1}$$

所以 P(s|s)和 P(s|n)就可以评价一个说话人确认系统，P(s|s)和 P(s|n)分别称为正确接受和错误接受的概率，P(n|s)为错误拒绝率。将他们分别作为横坐标和纵坐标，再改变系统的判决阈值，就可以对说话人识别系统画出 ROC 曲线（Receiver Operating Characteristic curve）。在图 1.5 中方法 B 始终优于方法 A，而 D 相当于没有识别能力的场合。

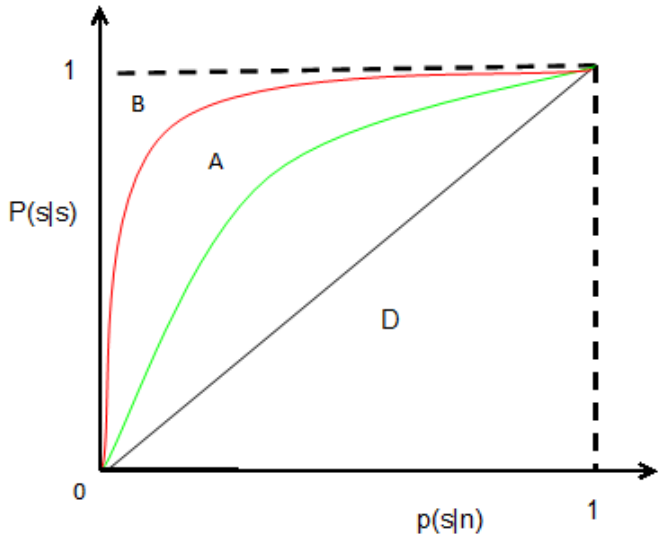


图1.5 ROC曲线

1.3.2 FA-FR 曲线和等误识率

在说话人确认中的错误率由错误拒绝概率（False Rejection，FR）

$$E_{miss} = \frac{n_{miss}}{n_{target}} \tag{1.2}$$

n_{target} 为目标说话人实验的次数，n_{miss} 为话者是目标说话人却未被检测出的次数。

和错误接受概率（False Acceptance，FA），

$$E_{fa} = \frac{n_{fa}}{n_{imposter}} \tag{1.3}$$

$n_{imposter}$ 为非目标人说话的次数，n_{fa} 是话者不是说话人却系统被检测认为是说话人的次数。

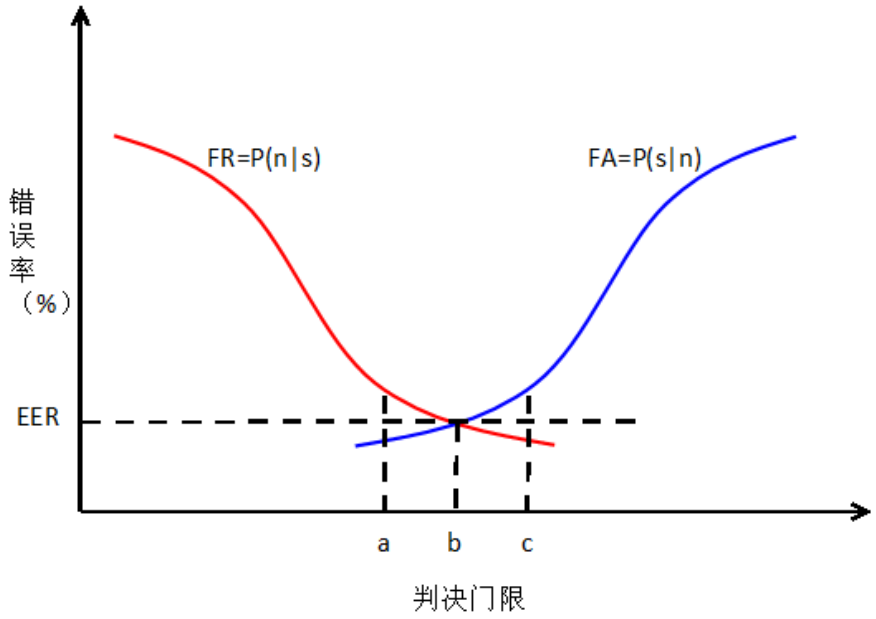


图1.6　两种判决门限的关系

也就是等同于上面介绍 ROC 曲线时的 P(n|s)和 P(s|n)。前者是拒绝真实的声言者而造成的错误，后者则是把冒名顶替者误认为其声言者而造成的错误在下面即将介绍的 DET 曲线上 FR 又称为漏检率（Miss probability），FA 又被称为虚警率（False Alarm probability）。

FR 和 FA 的曲线关系可见图 1.6，分别描绘了不同阈值确认下得到的 FR 和 FA，图中系统阈值为 X 轴，错误率（包括 FA 和 FR）为 Y 轴，阈值与 FA 呈反比关系（当阈值增加，FA 减少，此时冒认者越难接受，反之亦然），当然 FR 与 FA 成反比关系，即 FA 与阈值呈正比关系，而阈值的设定需考虑到 FR 和 FA 两种错误指标，图中明显指出了阈值与 FR 和 FA 之间的关系。在一些指定场所，应使 FA 值尽可能小，以免非指定人员的进入导致不可控的后果，为此，可将阈值指定在图中位置 c，此时 FR 应尽量大。反之，公共场所中有大量使用者时，数据需公开，对 FA 要求相对较低，此时阈值需做

相关调整，至图中 a 的位置[9]。等误识率 EER(Equal Error Rate)表示 FR 与 FA 相等的位置，即 FR 与 FA 两曲线的相交位置，见图中阈值 b 处，这是确认系统的很重要的指标，在常规使用时可采用此阈值，然而此阈值不能反映系统的识别性，那是由于需要考虑 FA 和 FR 对这个系统的产生的损坏程度。

1.3.3 检查代价函数

表 1.1 训练和测试段条件，阴影部分是必需的核心试验条件

		Test Segment Condition		
		10sec	**core**	**summed**
Training Condition	**10sec**	optional		
	core	optional	**required**	optional
	8conv	optional	optional	optional
	8summed		optional	optional

美国国家标准技术署（National Institute of Standards and Technology，NIST）在1996年以来每年一次的说话人识别评估（Speaker Recognition Evaluation，SRE）中定义了一个新的指标：检测代价函数（detection cost function，DCF)[3]。NIST对说话人识别尤其事对与文本无关的说话人确认的研究起到了不小的推动作用。

说话人辨认系统是多选一的问题，可以直接用错误率，也就是错误数/总测试数来表示；或者是正确率来表示，也就是正确数/总测试数。而对于说话人确认，系统将待识别语音与说话人模型进行匹配，然后得出测试评分。通过该评分与设定的阈值进行比较得到是否是真实说话人还是冒认者。

根据NIST的2010年的说话人识别评估计划(The NIST Year 2010 Speaker Recognition Evaluation Plan)[4]，训练和测试段条件在表1.1中，阴影框里的是2010的核心测试评价。

每个测试的审判结果必须各自独立的被判断为真确或错误，再对系统的正确性进行判断。检测函数被定义为加权的错误接受率（False Alarm probability）和错误拒绝率（Miss probability）之和（虽然和上面的表示有所区别，但是所代表的意义是一样的）：

$$\begin{aligned} C_{Det} &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &\quad + C_{FalseAlarm} \times P_{FlaseAlarm|NonTarget} \times (1 - P_{Target}) \end{aligned} \tag{1.4}$$

公式里的参数，P_{Target} 是目标说话人的先验概率，C_{Miss} 和 $C_{FalseAlarm}$ 分别为错误拒绝和错

误接受的代价。主要情况下取值为 C_{Miss}=10，$C_{FalseAlarm}$=1，P_{Target}=0.01。相对于等错误概率，由于不同的错误概率所带来的不同程度的代价和不同的先验概率，在实际应用中要更加科学合理，而且可以找出最小检测代价作为系统的一个评估手段，称之为最小DCF。可以用最小 DCF 表示系统能够取得的最优性能，最小 DCF 越小，系统的性能越好，从最小 DCF 还可以看出系统的侧重点是“安全性”还是“方便性”。在计算最小DCF 时候，如果不特别说明，也是用同样的数值。

为了提高 C_{Det} 的直观意义，将他除以 $C_{Default}$ 实现归一化，$C_{Default}$ 不处理输入数据就可以得到（即在匹配的目标说话人时总是接受或总是拒绝，两者的代价都更低）：

$$C_{Default}=\min\begin{Bmatrix} C_{Miss}\times P_{Target}, \\ C_{FalseAlarm}\times(1-P_{Target}) \end{Bmatrix} \tag{1.5}$$

和

$$C_{Norm}=C_{Det}/C_{Default} \tag{1.6}$$

这个数值能反映系统对一个语音片段来自目标说话人的语音片段的可能性的估计，数值越高意味着当前语音是目标说话人的语音概率就越高。可用来也可以用来描述DET曲线，以便能观察出错误拒绝率和错误接受的对比关系，因为曲线可以描述所有试验中每个说目标说话人测得结果，所以对数据进行归一化是有必要的。

1.3.4 DET 曲线

检测错误折中(Detection Error Trade-off)曲线[11]常用来描述一个说话人识别系统的性能，如图，是用不同确认阈值下得到的FA和FR的值而描绘而成。它也就是上面所提到的NIST 说话人识别评测中的检测标准曲线。DET 图根据FA 和FR 相应的高斯偏离程度，取对数形式，这种做法将导致非线性的概率尺度，但优点是该曲线将更为直观。尤其是当错误概率的分布服从高斯分布时，这是得到的检测错误折中曲线将为一条直线，曲线之间的距离大小将更有效地描述不同系统之间性能差异的大小[12]。由于NIST评测比赛中，需要将各个参赛系统的曲线放在一张图上进行比较，所以DET 曲线更加直观合理。

由于在 DET 曲线上，DET 曲线上的一点就代表了在一确认阈值下的 FA 和 FR，代表了说话人确认系统的确认性能，曲线越接近原点，系统的识别性能越好。可以从图1.7 中看出在黑色实线 a 所代表的识别系统性能最好。其中，图中在坐标 45°上虚线与曲线的交点也就是等误识率(EER)点。选取 EER 上方所选取的阈值，系统安全性要求较高，对说话人判别的 FA 值就小，以免非指定人员的进入导致不可控的后果，但是要牺牲 FR，也就是系统要求很严格，对说话人的语音的匹配率要求较高，反之，像公共场所或者电

话远程服务等即使被误接受也不会造成太大的损失时，对 FA 要求相对较低，就可以选取下方的阈值。对应于已经选定的检测函数，检查代价越小，则系统性能越好。图中先上面的点就是最小代价函数(min DFC)。由于 DET 曲线可以详细的描述系统的确认性能，可以刻画系统内在的整体分类性能，从图片中可以看出，不同系统可以很直接比较，效果也很直观，在说话人确认系统中得到了广泛的应用。

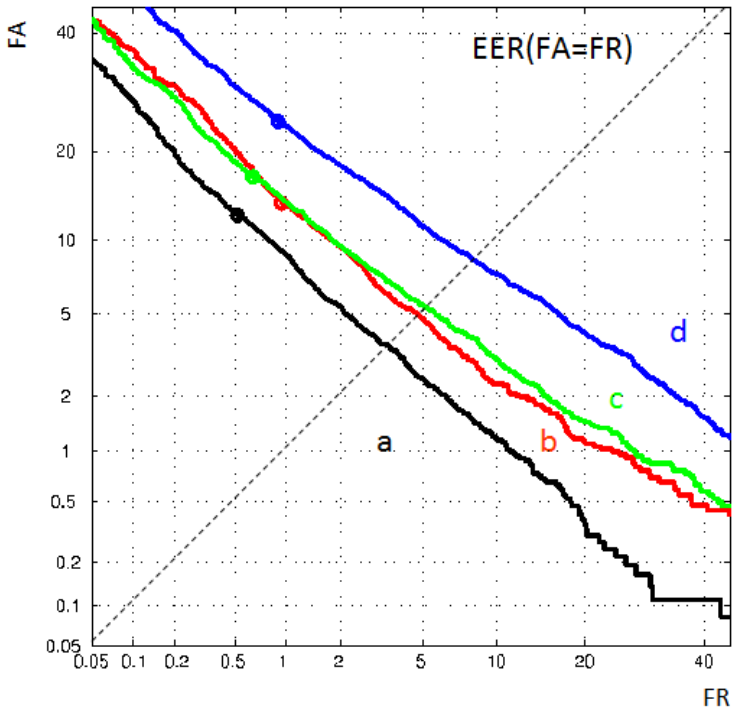


图1.7 DET曲线

1.4 复杂环境下的说话人识别

近年来说话人识别技术研究的逐步深入，虽然目前在限定条件下说话人识别已经可获得较为满意的识别效果，但是与实际应用的要求仍有一定距离，尚有一些问题亟待解决，说话人识别系统在实际应用中需要解决的一个关键问题是模型训练和应用环境的不匹配。在目前的使用环境下，造成这种不匹配主要有三种因素：背景噪音、传输信道和说话人的情感。这三类问题对说话人识别的影响都非常显著，本课题将针对背景噪声问题进行深入研究，集中在文本无关的说话人确认并分析说话人确认系统的鲁棒性。

鲁棒性算法的出现，其主要目的是为了增加识别各个参数的不同性能，其涉及的领域广泛，包括特征、分数、模型等等。相对于其在特征级上的应用，前部分可以通过相关方法来削弱因截断处理引起的Gibbs现象，比如窗函数，在此同时，可运用高频预来加重提高信息；也可以增强信号的语音部分，此种增强语音的方式是对噪声进行提取，

尽量恢复之前的无杂音的语音信息。通过这种方法得到的纯净语音信号，在前部分解决以及去除噪音的负面作用，增加识别性能，在众多的方法中，最普遍的方式为SS(Spectral Subtraction)，也称为谱减法。后段可运用倒谱的差分和ARC(Auto-Regression Coefficient)在非动态的倒谱中导入一些动态的信号来加强相邻帧之间的特征数据之间的共性。倒谱均值减与倒谱方差归一化通过去除整部分语音信息的倒谱均值，去除卷性信道对其的作用，特征弯曲与特征高斯化是在特征中增加短时特性来增强特征值的鲁棒性；RASTA亦被用作去除信道扭曲和噪声加性，进而对特征的各个维度在统计特征上做最优的归一化作用。

正如我们所熟知，一个或者多个语音音质的优劣，会直接作用于说话人识别体系的性能。但是在实际的运用情况中，因说话者处的环境千变万化，进一步增加了语音录制的困难，环境噪声无法避免的增加，不同的噪音混杂其中，如果不会对噪音做处理，系统的识别性能会随着信噪比的下降而急剧的下降。故噪音鲁棒性的问题一直是说话人识别研究的热点，也是其中的巨大难点之一。通常来讲，噪声所引起的失配可影响到特征、模型和信号三个空间和维度。

1.4.1 基于信号特征的噪音鲁棒算法

这种噪音鲁棒算法目的是为了消除含噪语音信号中的噪音成分，减少测试语音和训练语音之间因噪音带来的不匹配。该类算法主要是从信号处理的角度出发，或者去除噪音的影响，或者提高特征对噪音的抗干扰性。它包含了语音信号检测、噪音消除、信噪分离以及语音信号增强等众多技术。现有的方法很多：谱减(Spectral Subtraction, SS)法[27]、倒谱均值减(Cepstrum Mean Subtraction, CMS)、倒谱方差归一(Cepstrum Variance Normalization, CVN)[28]、特征弯折(Feature Warping, FW)[29]、相对谱(RelAtive SpecTrAl, RASTA)滤波[30]、特征映射(Feature Mapping)[31]等谐波分析、主分量分析(Principal Component Analysis, PCA)法[32]和线性判别分析(Linear Discriminant Analysis，LDA)[33]等。另外，多种特征进行融合，也可以取得较好的识别性能。这类算法的优点在于由于作用在特征域，可以通用于各种不同的说话人模型建模方式的说话人识别系统。

1.4.2 基于模型的噪音鲁棒算法

这种噪音鲁棒算法着眼于调整统计模型的参数，该类方法主要是在声学模型级上研究噪音问题，或采用模型模型补偿(Model Compensation)技术[34, 35]，这类算法的优点在于利用了语音和噪音的先验知识，针对不同说话人模型建模方式的特点组织、用先验的开发级数据，对语音模型进行补偿，增加测试集和训练集匹配程度，从而提高系统对含噪

语音的识别性能，例如因子分析(Factor Analysis, FA)等；或采用规避的方式建立不受噪声作用影响的说话人模型，这一类的代表算法为建立在高斯混合模型超向量-支持向量机模型(GMM Supervector SVM)的系统上的有害分量投影(Nuisance Attribute Project, NAP)。还有并行模型合并(Parrallel ModelCombination, PMC)[35]、加权投影法(Weighted Projection Measure, WPM)[36]等。

这几种方法在一定程度上都需要一定的噪声先验知识，而在实际运用时常常不能预知噪声种类。没有准确的噪声特征估计，这些方法的优越性就没法较好地得到体现。而抗噪声的语音特征提取能较好地弥补这个缺陷，因为抗噪声的特征参数不用预先知道噪声特征，而是通过寻找一种对噪声影响不敏感的语音特征参数来提高说话人识别性能，因此能适用于各种噪声环境。但是这种特征参数寻找比较困难，效果也不是很理想。现有的抗噪的语音特征参数大部分都是以上面所提到的三个特征参数为基础的。

1.5 本文使用到的语料库

1.5.1 TIMIT 语音库

语音数据库TIMIT是由LDC发布的第一个存在大量说话者的可用的语音数据库，从而得到了广泛的应用[8]。我们使用的是纯净的语音，然后在人工加上背景噪声，背景噪声的来自NOISEX-92数据库。

1.5.2 NOISEX-92 数据库

为了研究在噪声环境下识别系统的性能，我们还需要噪声的数据库作为辅助的数据库，背景噪音来源于NOISEX-92数据库，包含的噪声有：

- 白噪声（White noise）
- 粉红噪声（Pink noise）
- 高频无线电信道噪声（HF channel noise）
- 迷惑噪声（Speech babble）
- 工厂噪声（Factory floor noise 1，Factory floor noise 2）
- 飞机座舱噪声（Jet cockpit noise 1，Jet cockpit noise 2）
- 驱逐舰机舱噪声（Destroyer engine room noise）
- 驱逐舰作战室噪声（Destroyer operations room noise）
- 战机座舱噪声（F-16 cockpit noise）
- 军用车辆噪音（Military vehicle noise）

- 坦克噪声（ Tank noise）
- 机关枪噪声（ Machine gun noise）
- 汽车噪声（Car interior noise）

1.6 论文的安排

论文中首先对复杂环境下的说话人识别做一个大概的介绍，第2章中分析了语音信号的特征参数和3种常用的说话人识别的特征参数的提取。针对于在复杂环境的语音识别在第3章分析并仿真了常用的噪声鲁棒性的参数。在第4章中介绍说话人确认的模型，首先介绍说话人识别的高斯混合模型（GMM），并提出一个抗噪特征参数的提取方法。再将UBM引入到GMM中来应用于说话人确认，提高了与文本无关特定说话人确认模型对噪声的鲁棒性。最后介绍了SVM模型和SVM-GMM的混合模型并对所介绍到的几种模型的性能进行比较。

2 说话人特征参数提取

说话人识别特征参数提取的优劣能直接强烈的影响到分类器的设计及其性能。说话人识别过程中最重要的就是从语音信号中提取能表征说话人的基本特征。从语音感知的研究中可以看出不同说话人之间的差异包含了先天和后天的差异，这些差异都会不同程度上影响说话人识别系统的差异，也就是说话人之间的差异(inter-speaker)，而感情、健康等情况也会影响到说话人的发音情况，视为说话人本身（intra-speaker）的一些差异。说话人识别的特征参数就是要力求抑制intra-speaker的影响因素而突出inter-speaker之间的差异。而且一小段语音也包含了很大的数据量，它除了包含我们所需要的说话人个性特征的参数，还包含了语音识别所需要的语义特征，他们以很复杂的形式互相交织在一起。研究语音特征的目的就是寻找信号的最佳表示，以最少的特征数得到最好的识别性能。在说话人识别过程中，语音信号经过预处理，产生了大量的信息，但说话人识别只需要与具有能代表说话人特征的信息。说话人识别特征参数的提取就是要去掉原来语音中的冗余信息，减小数据量。

发音器官以及语音信号从整体上看是随着时间变化的，所以语音是一个非平稳态过程。不过虽然语音信号具有时变的特性，但通常可以被看为一个短时平稳的序列，在一个短时间内是可以被视为相对稳定的。因此，对语音信号进行分析和特征参数的提取一般是分为一段一段来进行的，也就是分帧处理，并利用分帧窗函数来减少由阶段处理导致的Gibbs效应。为了保证语音帧之间的连贯性，帧与帧之间须有交迭，也就是帧移。通过提升高频信息来压缩语音的动态范围。根据不同的识别任务和提取方法，可以得到许多种不同的语音特征参数。下面仿真图里面的语音都是用的TIMIT语音库中的纯净语音，语音内容为“She had your dark suit in greasy wash water all year.”

2.1 语音信号的预处理

2.1.1 预加重（Pre-emphasis）

将语音信号通过一个高通滤波器：

$$H(z)=1-\alpha z^{-1} \tag{2.1}$$

式中α介于0.9和1.0之间。

这个过程是为了消除发声过程中得口唇辐射的影响，补偿语音信号收到发音系统所压抑的高频部分。图2.1 所示的就是对语音进行预加重的结果图。

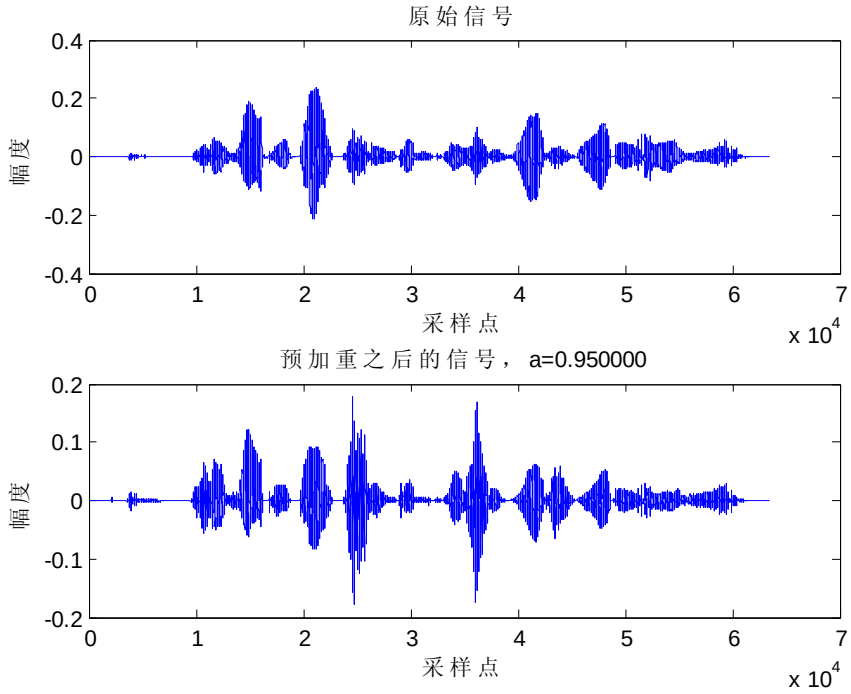


图2.1 原始语音和预加重之后语音的比较

2.1.2 加窗处理（Frame Blocking）

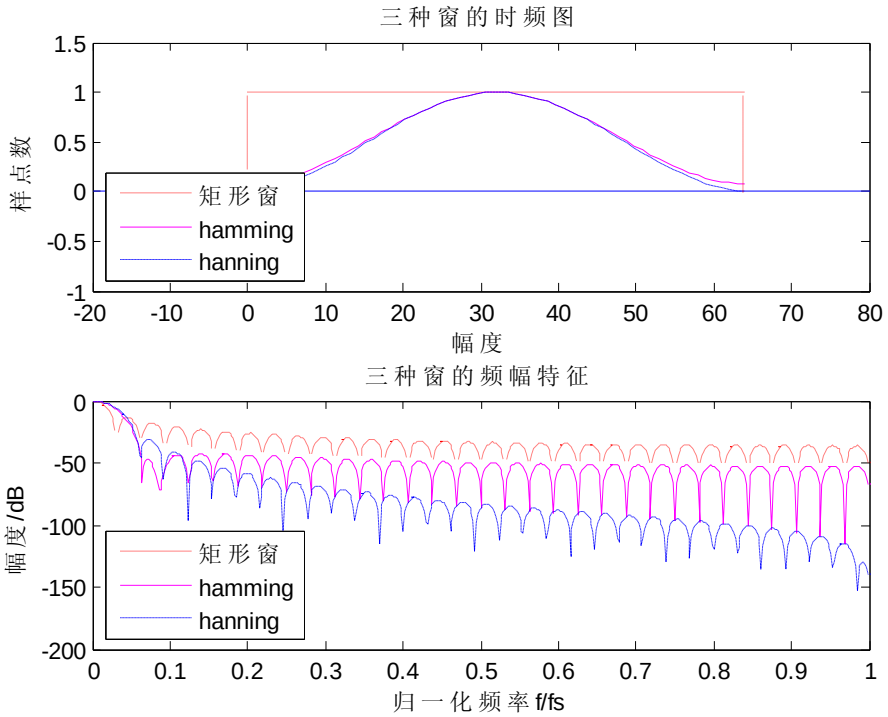


图 2.2 三种窗的时频和幅频特性的比较

进行预加重滤波处理之后，接下来就是进行加窗分帧处理。加窗的基本思想是，由于发音器官的的惯性运动，一般认为语音信号在一小段时间里是近似不变的，即语音信号的短时平稳性。所以需要加窗来实现对语音信号的分帧，一般情况下每帧的长度为 20ms 至 40ms，帧移为帧长的 1/2 或者 1/3。分帧所用的窗一般为三种：矩形窗、汉明窗和汉宁窗。我们取窗长为 64，三种窗的时域和频域的对比图如 2.2 所示。从在时频图中可以看出汉宁窗比汉明窗在边缘时更加趋近于零。在幅频特征中汉明窗和汉宁窗的主瓣宽度相同，矩形窗的主瓣宽度要小于前两者，具有较高的频谱分辨率，但是矩形窗的旁瓣值较大，因此频率泄漏比较严重。汉明窗旁瓣衰减较大，具有更平滑的低通特性。窗函数的要求就是主瓣越窄越好，边瓣越小并且衰减越快越好，一般情况下选用的窗为汉明窗来减轻吉布斯效应（Gibbs）。图 2.3 显示了在 α 取值不同时的汉明窗图。

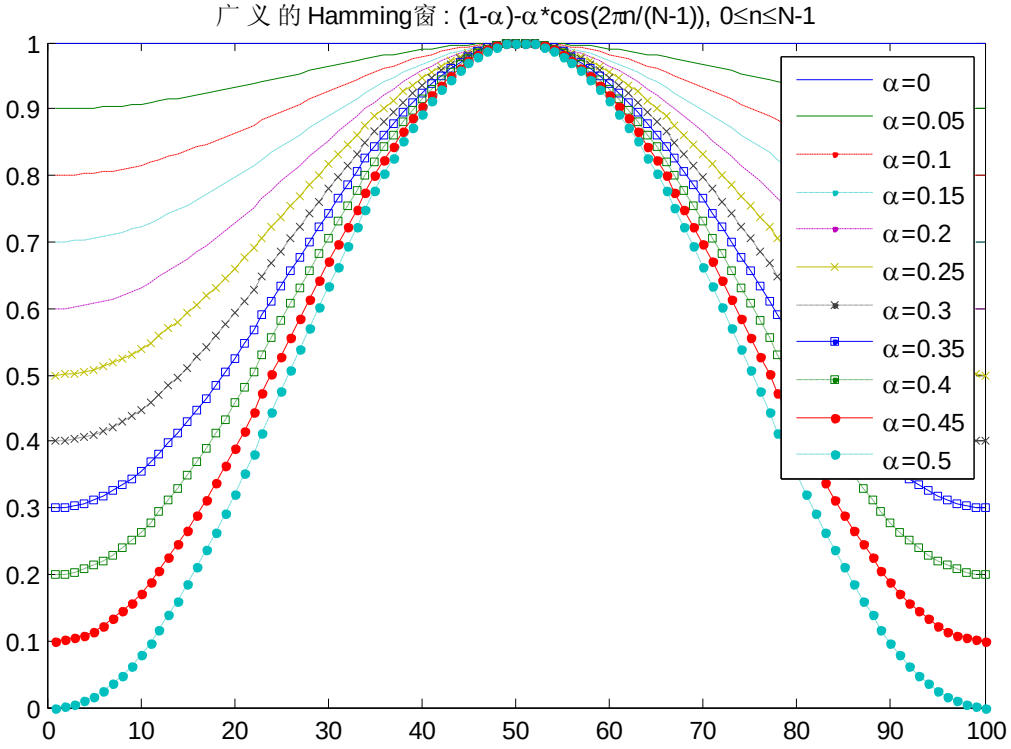


图2.3 汉明窗

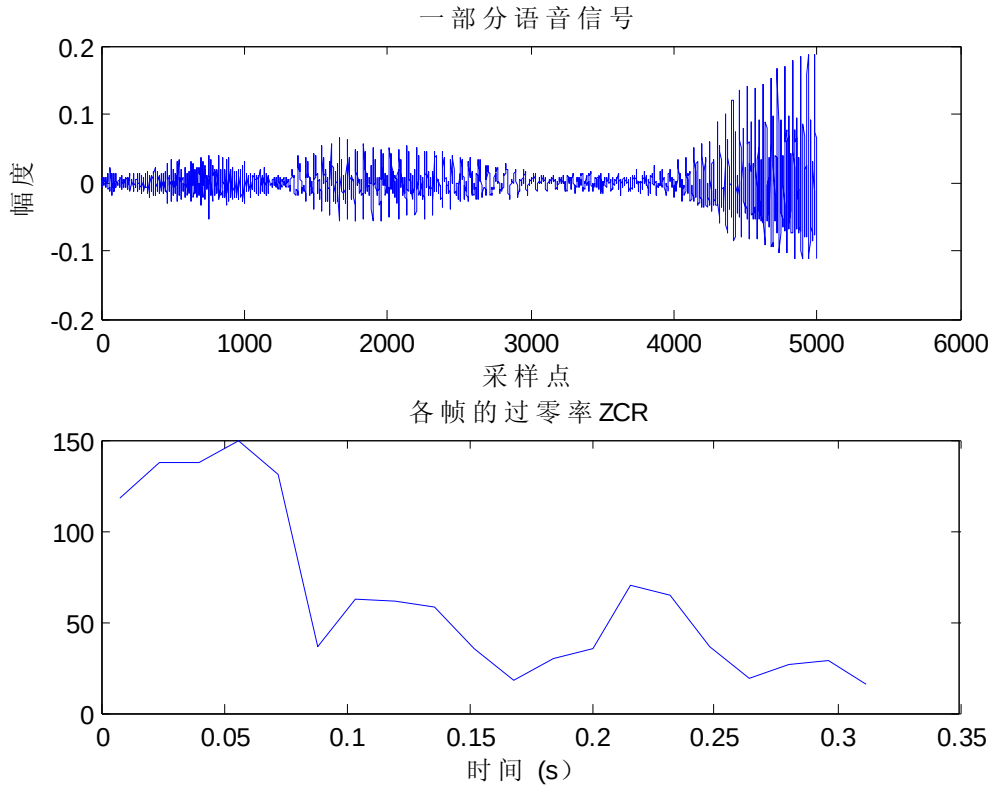
2.2 语音信号的特征参数的分类

特征参数这里主要分为下面的四类[14]来介绍。

2.2.1 时域特征。

直接从时域信号计算得到，反应了语音信号时域波形的特征。如短时平均能量、短时过零率、共振峰、基音周期等。在这里简单介绍一下。

（1） 过零率（Zero Crossing Rate, ZCR）指每帧信号通过零的次数，可用于端点检测。图2.4所示的是一段语音的过零率，帧长选用的256。



2.4 过零率

（2） 共振峰。

共振峰是喉管口腔鼻腔以及舌头腮帮子共同组成的发音系统固有的谐振频率 。共振峰指在声音的频谱中能量相对集中的一些区域，共振峰不但是音质的决定因素，而且反映了声道（共振腔）的物理特征。一般来说，频谱能量的个数很多，因此采用共振峰这个频谱能量图的局部最大点，来作为语音的特征，图2.5在语音信号中截取了信号的一个基本周期（样本数为64，其余的部分补零，本例中是手动截取的，可以用额外的算法先算出准确的值），然后算出频谱能量和对应的共振峰。

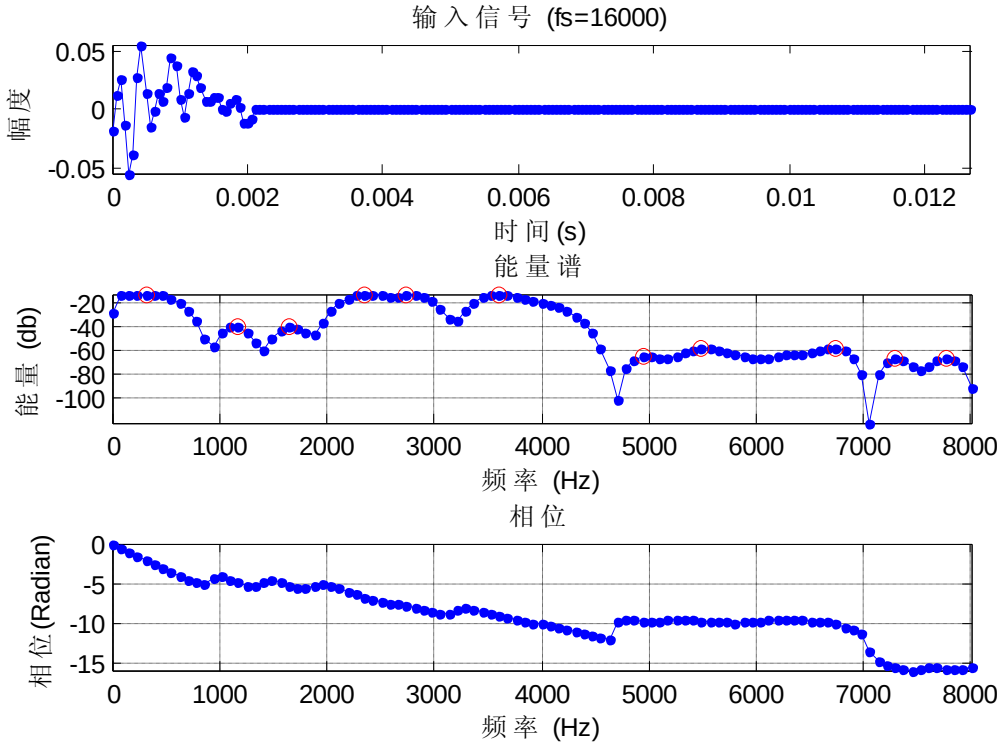


图2.5 共振峰

（3） 基音周期。基音（Pitch）是表征声带震动频率的参数，基音周期就是声带振动频率的倒数。

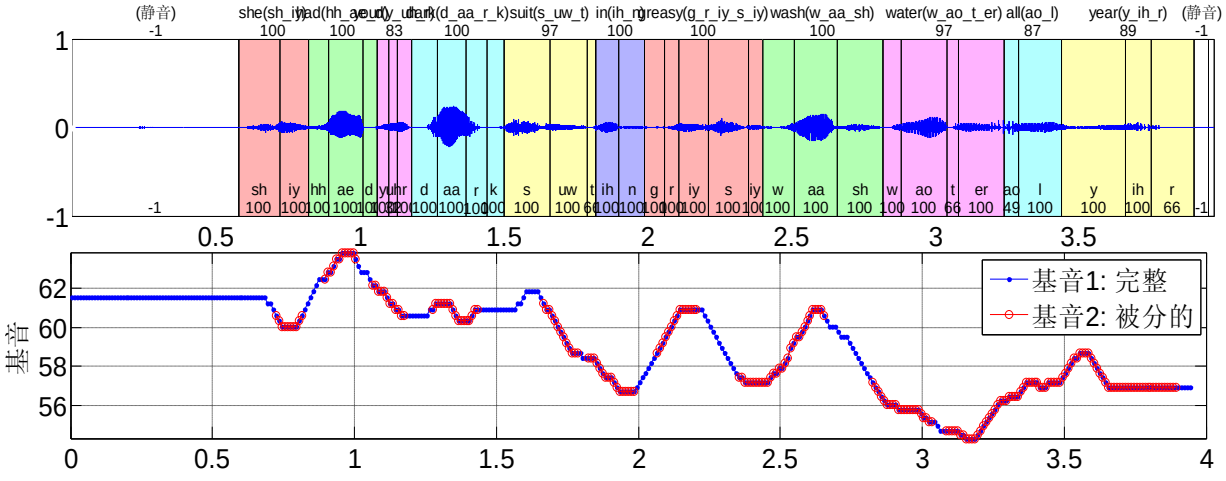


图 2.6 语音的基音

基音周期是语音产生的效字模型中激励源的一个重要参数，携带着非常重要的辨意信息，主要用在语音识别。基音周期检测的方法很多，主要分为3类：时域估计、变换域法和混合法。图2.6所示为使用ASR工具箱找出的语音的基音。图2.7是对信号的其中

一帧检测基音周期。

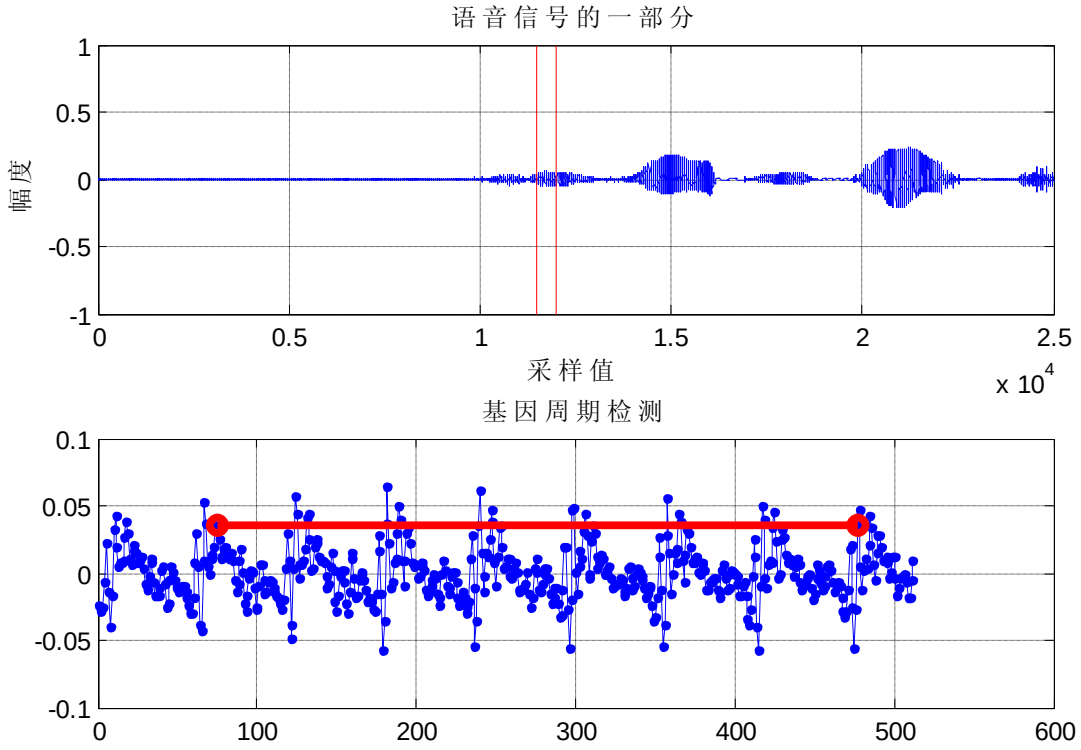


图2.7 基音周期的检测

2.2.2 线性预测编码参数。

以线性预测导出的各种参数，可以得到较好的效果，其主要原因是线性预测同声道参数模型不谋而合。

线性预测（Linear Prediction）的基本思想是：由于语音信号样点之间存在相关性，可以用过去的样点值来预测现在或者未来的样点值，即一个语音的抽样能够用过去若干个或者他们的线性组合来逼近[1]。

图 2.8 为线性预测参数（LPC）为语音其中一帧信号的估计信号和这个估计信号与原始信号的误差对比图。线性预测参数有很多的通过 LPC 分析，由若干帧语音可以得到若干组 LPC 参数，每组参数形成一个描绘该帧语音特征的矢量，即 LPC 特征矢量。由 LPC 特征矢量可以进一步得到很多种派生特征矢量，例如线性预测倒谱系数（LPCC）、线谱对特征（LSP）等。不同的特征矢量具有不同的特点，它们在识别领域有着不同的应用价值。在说话人识别中较常用到的是 LPCC，下面会单独介绍。

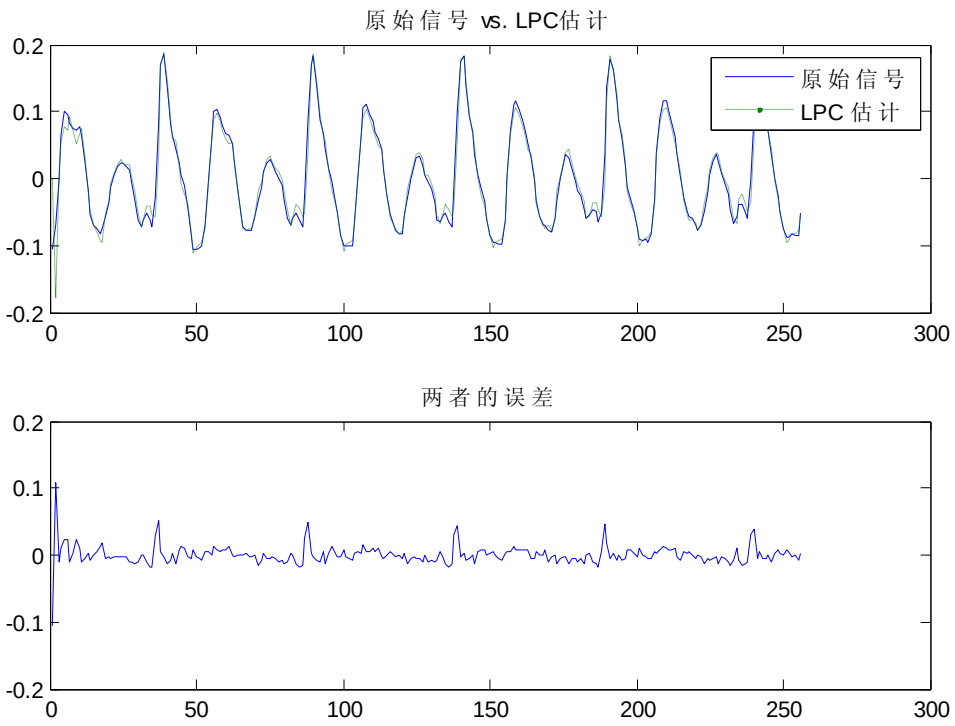
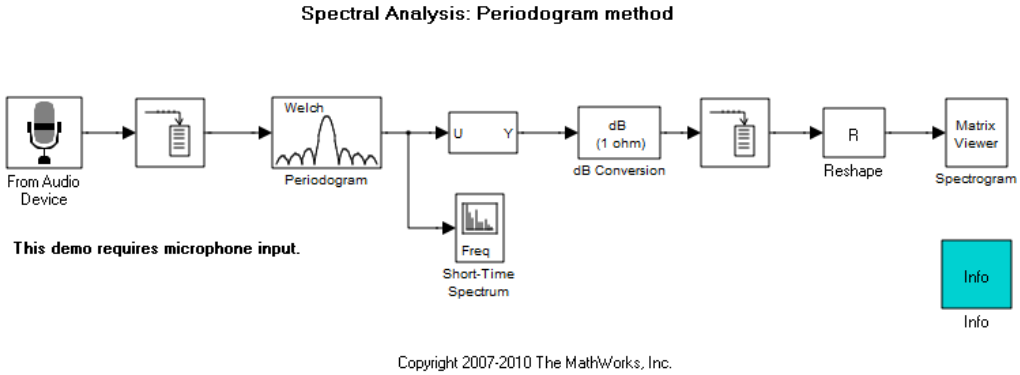


图2.8 语音信号的LPC估计

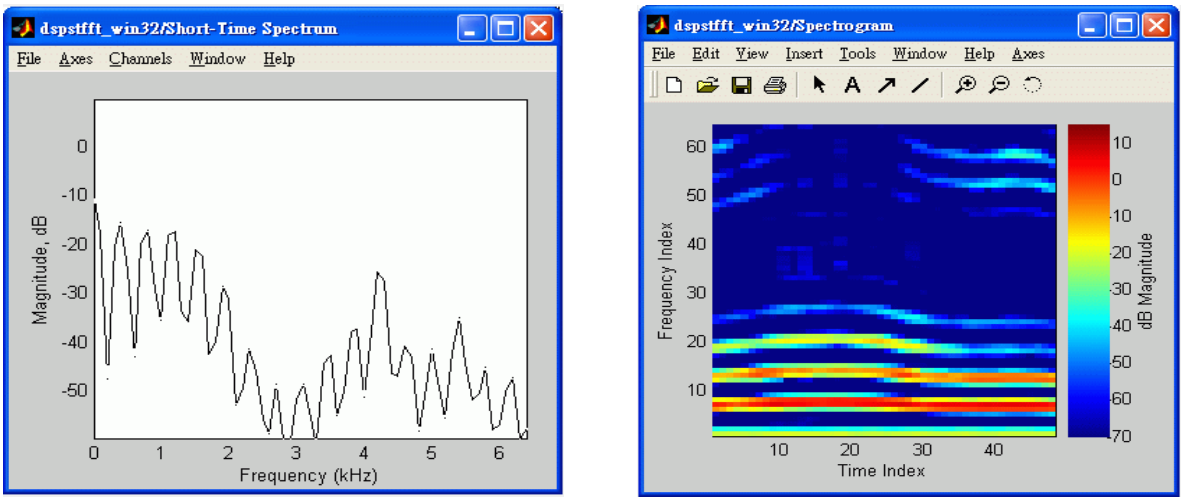
2.2.3 频域及倒谱特征。

由时域信号进行频谱变换得到，反映语音信号的频域特性。包括傅里叶频谱、倒谱以及利用了语音信号的时序信息的时频谱等。

（1） 傅里叶变换，若要从基本周期这些时域的特征来直接分析语音的音色，是比较困难的。通常需要对语音进行频谱分析（Spectral Analysis），在频谱分析时，最常用的就是快速傅里叶变换（FFT）。对语音matlab中傅里叶变换的Simulink系统如图2.9所示。



2.9 Simulink在的FFT变换系统图



（a）FFT频谱图　　（b）FFT谱图

图2.10 Simulink的FFT系统图所产生的频谱图和谱图

在启动程式之后对麦克风说话时候，就出现动态的频谱图和频谱对时间产生的影响的谱图Spectrogram。它代表了音色随时间变化的资料。如图2.10所示

对一长段的语音信号需要分帧之后再进行傅里叶变换。依旧用TIMIT中的语音，对其中的一帧进行加窗之后进行傅里叶变换，图2.11所示的是一帧语音信号的傅里叶变换。

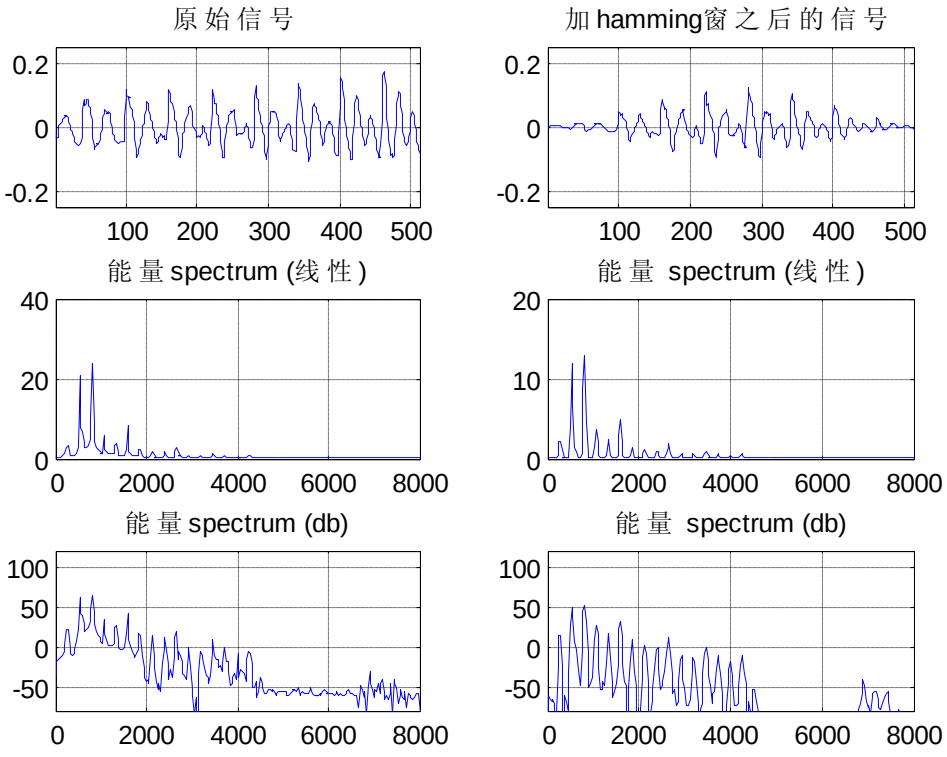


图2.11 语音信号的傅里叶变换

从图2.11中还可以看出加汉明窗对信号FFT变换的影响，它可以加强了各个语音帧左边和右边的连续性。在进行FFT变换时，假设的前提是一个帧之类的语音信号代表的是一个周期性的信号，如果这个周期性不存在，FFT会为了符合左右端不连续的变化而产生一些原信号中并不存在的能量分布，造成分析上的误差。如果在分帧时窗长取的是基本周期的整倍数，每帧左右是连续的，就不会产生这样的误差，也不需要加上汉明窗。但是在实际操作，由于基本周期的计算需要额外的时间且容易出错，要满足上述的条件是很困难的，因此都用汉明窗达到类似的效果。在带有噪声的语音信号中使用得到的效果会更加明显。

（2） 倒谱（cpstrum），语音信号的倒谱分析是求取语音倒谱特征参数的过程。他可以作为一个求取语音信号的方法。求取倒谱特征有两种，一种是上面所介绍的线性预测分析，另一种的同态分析处理。同态信号处理也称为同态滤波，实现对卷积信号的解卷过程。一个信号的倒谱c(n)定义为信号s(n)频谱的自然对数的逆傅里叶变换：

$$\hat{s}(n) = IDFT\{\ln|DFT[s(n)]|\} \qquad (2.2)$$

图2.12现实了对一帧语音信号进行傅里叶变换之后所求取的倒谱。图中圆圈所圈出的基频的位置，倒谱域中基音信息和声道信息可以被认为是相对分离的，就可以采取简单的倒滤波方法就可以分离并恢复两者。

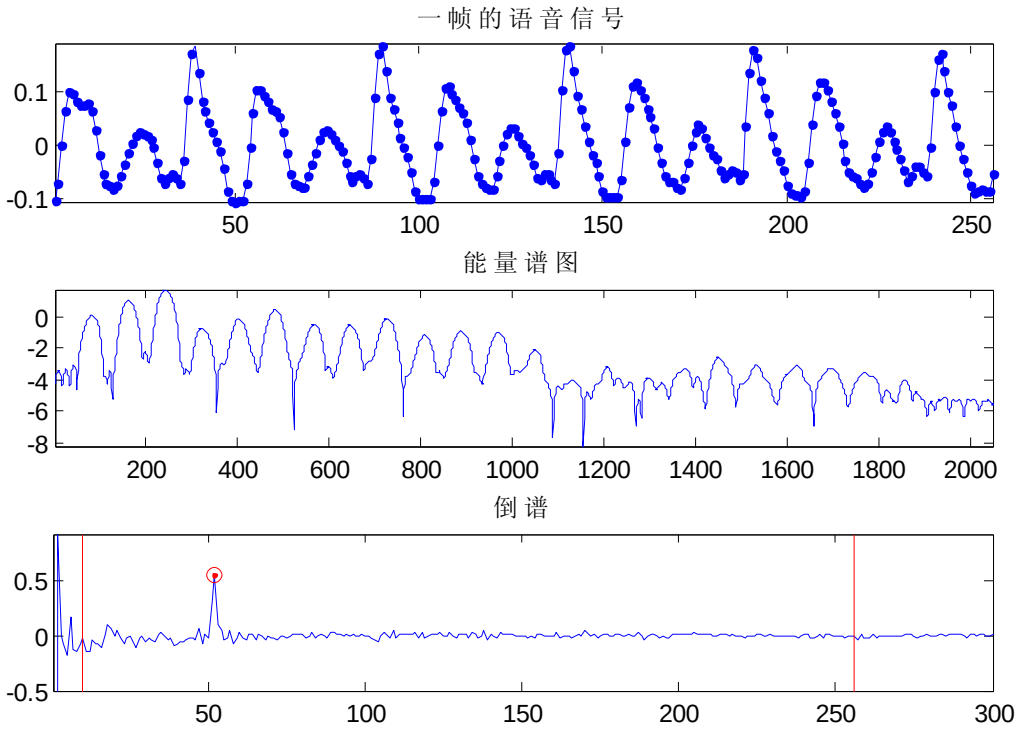


图2.12 语音信号的倒谱

2.2.4 基于听觉特征的参数。

基于听觉特征的参数不是传统的直接对声道模型进行研究，而是模拟人耳对声音的感知特性来刻画语音信号的特征。常用到是MFCC和PLP，会在下面的章节中进行介绍。由于这种特征参数不依赖于信号的性质，对输入信号不作任何的假设与限制，又采用了听觉模型的研究成果，能具有较好的鲁棒性。

目前常用的特征参数有根据语音信号的全极点模型得到的线性预测倒谱参数LPCC、根据人耳对不同频率的语音信号的敏感程度提取，在提取参数时利用Bark 刻度对语音频谱经行了刻度转换，模拟人的听觉特性的Mel 频率倒谱系数MFCC 和感知线性预测参数PLP等等。

2.3 LPCC

倒谱是语音信号幅度谱对数的傅里叶变换，线性预测倒谱参数（Linear Prediction Cepstrum Coefficient, LPCC）是线性预测系数（Linear Prediction Cepstrum, LPC）在倒谱域中的表示[5]。能较好的提取语音信号的谱包络，反映出语音信号的声道特性。

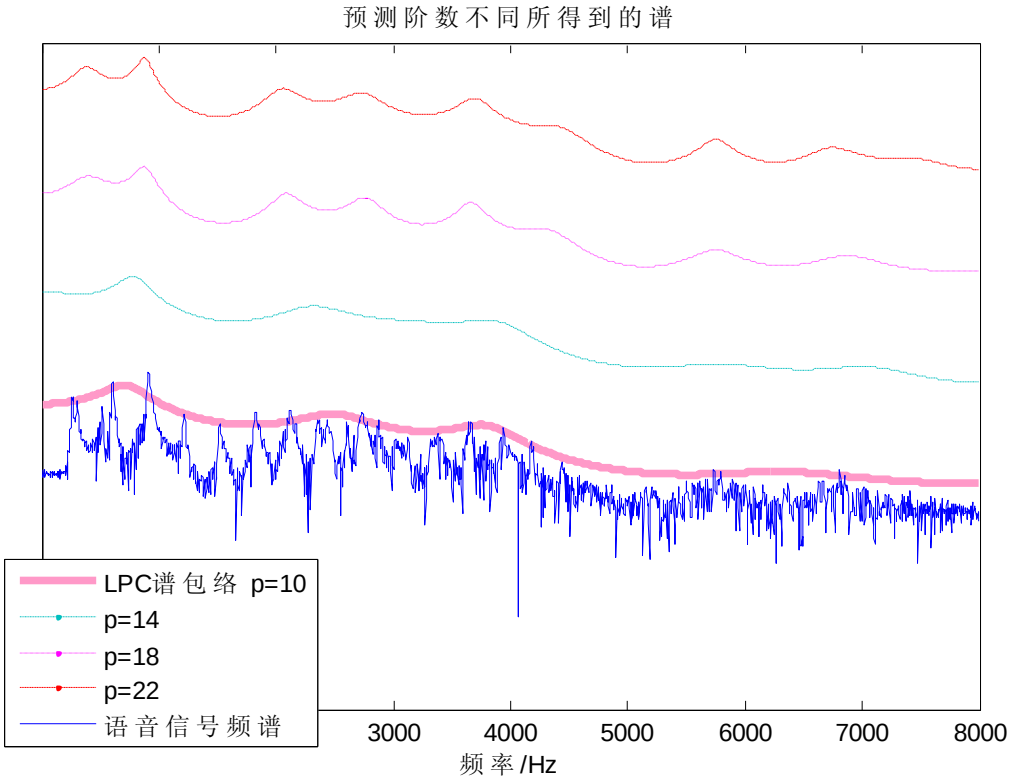


图 2.13 LPC 谱包络

LPC通过自相关法求得，具有稳定性。语音信号的倒谱c(n)与LPC系数之前的递推

关系为：

$$\begin{cases} c(1)=a_1 \\ c(n)=a_n+\sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c(n-k) & 1<n\le p \\ c(n)=\sum_{k=1}^{p}(1-\frac{k}{n})a_k c(n-k) & n>p \end{cases} \quad (2.3)$$

或由 LPC 得到

$$C_{LPCC}(n)=C_{LPC}(n)+\sum_{k=1}^{n-1}\frac{n-k}{n}C_{LPCC}(n-k)C_{LPC}(k) \quad (2.4)$$

声道传输主要携带了语音信号的语义信息，在语音识别中通常取语音信号倒谱的低频部分构成LPC倒谱特征c

$$c=[c(1),c(2),\cdots, c(q)] \quad 10\le q\le 16 \quad (2.5)$$

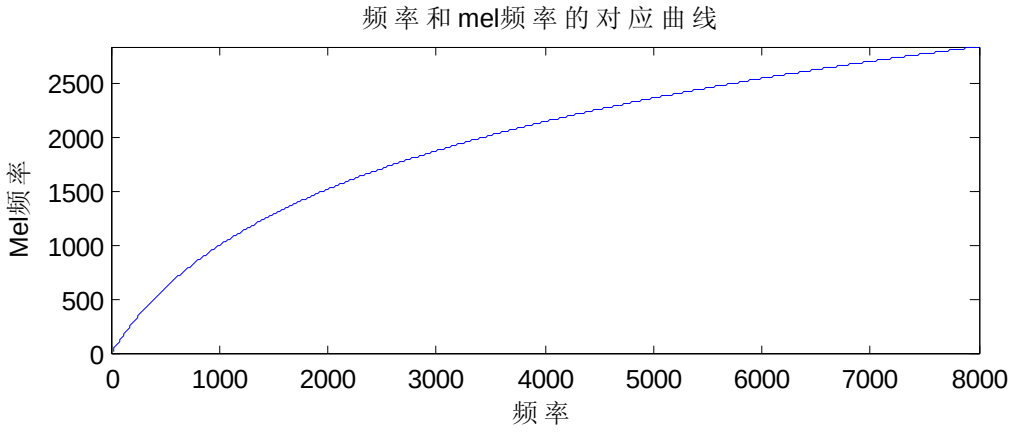
q 为 LPC 倒谱特征的阶数。C_0 反映频谱的能量，一般不使用。图 2.12 是以 16kHz 的采样的语音信号，在阶数不同时所得到的谱。可以看出随着阶数的增高，所得到的包络谱就会越接近原始语音信号的频谱，但到 p 大于 18 之后，对性能的提高就不再明显了。

2.4 MFCC 特征

Mel 频率倒谱系数（Mel Frequency Cepstrum Coefficient, MFCC）主要应用于文本无关的说话人识别。语音信号可视为声门激励信号和声道响应的卷积，因此语音信号的短时频谱是声门激励信号的频率响应与声道传输函数的乘积。一般认为声道传输函数与语意密切相关，而声门激励信号的频率响应则被视为与说话人信息相关。理想的情况是，提取声道传输函数作为特征来进行语意识别，而提取声门激励信号的频率响应来进行说话人识别。但是，实际情况是由于人的发声过程中声门激励信号和声道的作用并不是线性的，因此，要完全分离这两种信息是不现实的。我们只能近似将两者分离，这就是语音信号处理中常常提到的同态解卷，即倒谱分析[19]。

其基本思想是，用对数算子将声门激励信号的频率响应与声道传输函数的乘积变成各自对数频率响应的和，由于声门激励信号的频率响应从频谱上看是一个快变信号，而声道传输函数从频谱上看是一个慢变信号。因此，如果我们将频率轴视为时间轴，那么声门激励信号的频率响应的“频谱”应主要集中在“高频”区；而声道传输函数的“频谱”应主要集中在“低频”区。这样，我们就可以将两者分开了，线性频率f与Mel频率之间的转换关系为：

$$Mel(f)=2595\log_{10}(1+\frac{f}{700}) \quad (2.6)$$



HTK采用的是修正之后的公式：

$$F_{mel}=1127.0\ln_2(1+\frac{f}{700}) \tag{2.7}$$

Mel频率实际上是在有限的频率阈空间适当扩大低频部分分辨率，这正是人耳接收语音信号时的感知特性，但这是以牺牲高频部分分辨率为代价的。也就是说 Mel频率分辨率会随着频率的增加逐渐减小，这削弱了Mel频率域在高频部分不同频率带的频谱差异，对其后的特征参数产生不利影响。

2.4.1 MFCC 参数的具体计算过程

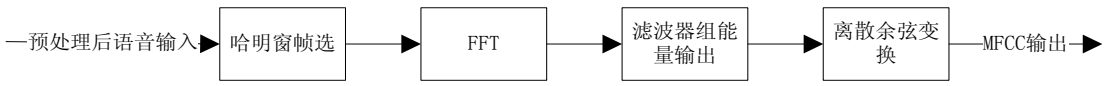


图2.1 语音MFCC参数的提取过程框图

（1） 对信号预处理：预加重，然后分帧（本文所使用的信号采用 16kHz 采样频率，16bit 量化精度。取帧长为 16ms，也就是 256 点为一帧，帧移为 8ms），并加 Hamming 窗。

$$w(n)=\begin{cases}0.54-0.46\cos\dfrac{2\pi n}{N-1} & 0\le n\le N-1\\ 0 & \text{其他}\end{cases} \tag{2.8}$$

$$H_w(n)=h(n)\times w(n) \tag{2.9}$$

w(n)是窗信号，$H_w(n)$是加窗后的信号。

（2） 对信号进行预处理之后，针对每一帧作短时傅里叶变换(Discrete Fourier Transform, DFT)。将时域信号转换为频域分量，得到语音信号的离散频谱。

（3） 按 Bark 刻度把频率分成 D 等分，计算中心频率和边界频率，然后根据它们求出频谱平方，即能量谱，用序列三角波滤波器进行滤波器进行滤波处理，得到一组系数 Y_1，Y_2，…,Y_D。计算公式如下：

$$Y_i = \sum_{k=F_{i-1}}^{F_i} \frac{k-F_{i-1}}{F_i-F_{i-1}} X_k + \sum_{k=F_{i+1}}^{F_{i+1}} \frac{F_{i+1}-k}{F_{i+1}-F_i} X_k \qquad i=1,2,\ldots, D \qquad (2.10)$$

其中，X_k 为频谱上第 k 个频率点得能量；Y_i 为第 i 个滤波器的输出；F_i 为第 i 个滤波器的中心频率。

（4） 将每个滤波器的输出取对数，得到相应频带的对数功率谱；并进行反离散余弦变换，也就是离散余弦变换（Discrete Cosine Transform，DCT），过渡到倒谱域，得到 L 个 MFCC 系数，一般取 12~16 个左右。MFCC 系数为：

$$C_k = \sum_{j=1}^{D} \log(Y_j) \cos \frac{k(j-0.5)n}{D} \qquad k=1,2,\ldots, p \qquad (2.11)$$

其中，p 为 MCFF 参数的阶数；{C_1，C_2，…，C_k}即为所求的 MFCC 参数。

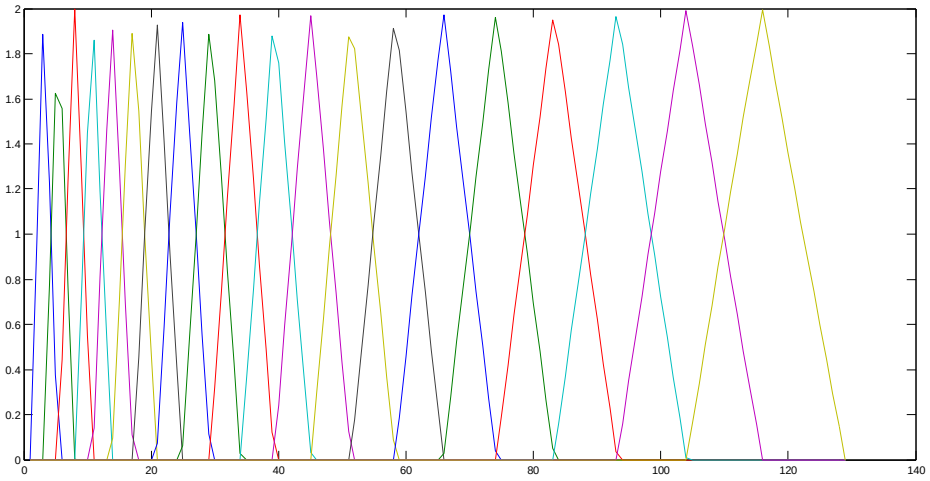


图2.2 MFCC滤波器组

根据文献[3]无论是在训练数据充足或者不充足的情况下，MFCC 阶数较低时，识别率较低，随着阶数的增加，识别率随之增加。但当阶数达到 16 附近时，识别率就不再增加了，甚至还出现了识别率降低的趋势。在求取 MFCC 参数时，用 D 个三角滤波器进行滤波，得到输出 Y_1，Y_2，…，Y_D，然后代入上式，得到 p 维的 MFCC 参数。

它将频谱转化为基于频率的非线性频谱，然而转换到到频谱域上，由于充分考虑了人耳的听觉特性，在一定程度上模拟了人耳对语音的处理特点，而且没有任何提前假设，MFCC 参数具有良好的识别性能和抗噪能力，在信道噪声和频谱失真的情况下具有较好的稳健性。但是计算量和计算精度要求较高。

2.4.2 临界带宽

这里介绍一下临界带宽的概念。噪声的存在会对纯音产生掩蔽，使纯音的听阈上升，一个纯音可以被以它为中心频率、并且具有一定频率带宽的连续噪音所掩蔽，若在这一频带内噪声功率等于该纯音的功率，则该纯音处于刚能被听到的临界状态，这一带宽称为临界带宽(Critical Band)[22]。临界带宽的单位用巴克（Bark）来表示。在20Hz~16kHz 的范围内可划分24 个Bark，当某纯音位于掩蔽声的临界频带之外时，掩蔽效应仍然存在。临界带的分布在1kHz 以下近似线性，而在1kHz 以上近似对数关系，临界带宽的增长与感知频率的增长是一致的。根据临界带的划分，将语音在频域上划分成一系列三角形的滤波器序列，即Mel 滤波器组，如图所示。用Mel 滤波器组对语音频谱进行滤波和加权，使语音信号更加逼近人耳的非线性听觉感知特性，这对于提取更加有效的语音特征参数具有积极的意义。图中滤波器的具体阶数应根据语音信号的采样频率而定，一般最高阶的滤波器的中心频率不应大于采样频率的一半。取每个临界带内经滤波加权后的信号的能量作为输出，再作对数运算，最后作离散余弦变换即得到Mel 倒谱参数(MFCC)。

2.5 PLP 参数

感觉线性预测参数(Perceptual Linear Predictive, PLP参数)是一种基于听觉模型的特征参数。等效于一种LPC 特性。不同的是LPC 特性使用的是时域信号，但是在PLP 特性中使用的是语音信号听觉模型处理之后的信号，更加利于语音信号的处理。

PLP 模仿人耳的听觉感知机理主要表现在三个方面：

（1） 临界频带分析处理（critical-band spectral resolution）

（2） 等响度曲线预加重（The equal-loudness pre-emphasis）

（3） 信号强度-听觉响度变换（The intensity-loudness power law of hearing）

2.5.1 PLP 特征提取的过程

语音信号经过采样量化，然后预处理和分帧。分别对每一帧进行 FFT，得到能量谱，再计算临界带听觉谱。临界带宽 Z(Bark)与频率 f(Hz)之间的关系为：

$$Z = 6 \times In(\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}) \qquad 0 \le f \le 8kHz \tag{2.12}$$

式中：

$$f = 600 \sin(B/6) \qquad 0 \le B \le 25 \tag{2.13}$$

根据上式，将P(f)的频率频率轴映射到Bark频率Z上，进行临界频带划分，再寻找每个临界频带的中心频率$Z_0(k)$。根据$Z_0(k)$可以求出加权系数

$$\Psi(Z-Z_0(k))=\begin{cases}0 & Z-Z_0(k)<-1.3\\ 10^{(z-z_0(k)+0.5)} & -1.3\le Z-Z_0(k)\le -0.5\\ 1 & -0.5<Z-Z_0(k)<0.5\\ 10^{-2.5(z-z_0(k)-0.5)} & 0.5\le Z-Z_0(k)\le 2.5\\ 0 & Z-Z_0(k)>2.5\end{cases} \tag{2.14}$$

再根据 $Z_0(k)$求出其所对应的频率 $f_0(k)$

$$f_0(k)=\frac{1}{2}\times 600\times(e^{\frac{Z_0(k)}{6}}-e^{-\frac{Z_0(k)-2.5}{6}}) \tag{2.15}$$

每个 Bark 中所对应的低端频率 $f_l(k)$和高端频率 $f_h(k)$为

$$f_l(k)=\frac{1}{2}\times 600\times(e^{\frac{Z_0(k)-2.5}{6}}-e^{-\frac{Z_0(k)-2.5}{6}}) \tag{2.16}$$

$$f_h(k)=\frac{1}{2}\times 600\times(e^{\frac{Z_0(k)+1.3}{6}}-e^{-\frac{Z_0(k)+1.3}{6}}) \tag{2.17}$$

式中$Z_0(k)$是第k个Bark的中心频率。

根据低端频率$f_l(k)$和高端频率$f_h(k)$，可得到每Bark所对应的最低点和最高点，就可以知道每个点的具体加权系数。将语音信号的短时功率谱与加权系数相乘，求和后后就可以得到临界带宽听觉谱θ(k)

$$\theta(k)=\sum_{N=n_l(k)}^{n_h(k)}p(f(N))\psi(Z(N)-Z_0(k))\qquad(k=1,2,\cdots,) \tag{2.18}$$

$N_l(k)$表示第k个临界带听觉谱对应的低端点，$N_h(k)$则表示最高点。

用模拟人耳大约40dB等响曲线E(f)对θ(k)进行等响度曲线预加重

$$\Gamma(k)=E[f_0(k)]\bullet\theta(k)\qquad(k=1,2,\cdots,) \tag{2.19}$$

$f_0(k)$表示第k个临界带听觉谱的中心频率，在临界带分析时可以求得

$$E[f_0(k)]=\frac{(f_0^{\,2}(k)+1.44\times10^6)f_0^{\,4}(k)}{(f_0^{\,2}(k)+1.6\times10^5)^2\times(f_0^{\,2}(k)+9.61\times10^9)} \tag{2.20}$$

是等响度预加重的权值系数，近似地反映了人耳对不同频率的不同敏感性。

为了近似和模拟声音的强度与人耳感受的响度间的非线性关系，在等响度曲线预加重之后进行响度幅值的压缩

$$\Phi(k)=\Gamma(k)^{0.33} \tag{2.21}$$

之后用全极点模型求线性预测系数，最后经行倒谱计算。

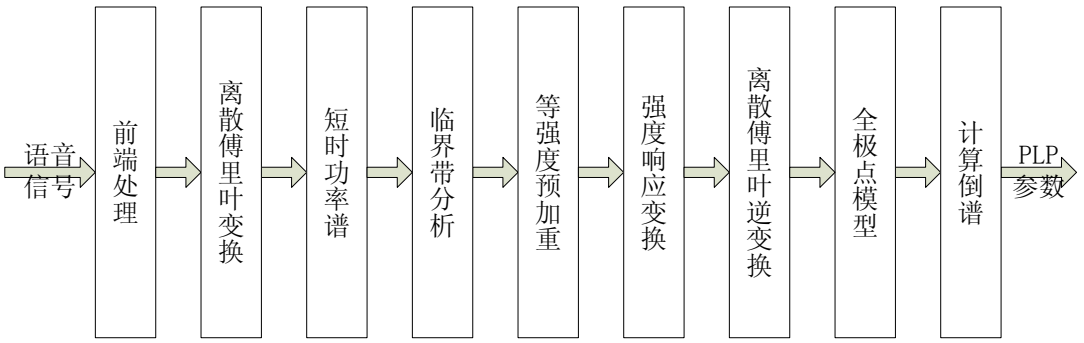


图2.3 语音信号PLP参数提取过程框图

2.6 小结

本章详细介绍了 LPCC、MFCC 和 PLP 三种常用的特征提取算法，和提取参数的过程。三中参数的各个特点不同，PLP 和 MFCC 在说话人识别中要比 LPCC 好。据 Reynolds 的研究表明[21]，在说话人识别任务中，MFCC 比 LPCC 和 PLP 具有更优越的识别性能。相对与 MFCC，LPCC 参数的优点是计算量较小，容易实现，对元音有较好的描述能力，但是对辅音的描述能力较差，抗噪性能也不是很好。LPCC 同样继承了 LPC 的缺陷，一般都不建议使用单独的 LPC 特征，经过大量实验，MFCC 特征的识别率要比 LPC 特征的识别率高出 7%。LPC 在所有的频率上都是线性逼近语音的，与人的听觉特性不一致；而且 LPCC 反映了语音高频系数的噪声细节，直接影响到系统的性能。

3 说话人识别中特征提取的鲁棒性

对于低噪声环境下的语音，目前说话人识别已能达到较高识别率，但在噪声的系统中系统性能会随着噪声信噪比的提高很大程度的降低，而在现实生活中，噪声是无法完全避免的，所以想办法提高说话人识别算法的噪声鲁棒性至关重要。为了提高系统的识别性能，提出了各中对特征参数的线性和非线性变换方法。为了在静态的每帧的倒谱中加入各帧之间的动态信息提出倒谱的差分（Difference，也称为Delta参数），针对信号的“高维灾难”提出了PCA，通过减去噪声能量谱而净化语音的谱减法和非线性的谱减法等等。

3.1 Delta 参数

一般语音特征的提取都是假设语音信号每帧之间是相互独立的，只考虑到语音帧内（intra-frame），却没有考虑到语音帧间的关系（inter-frame）。而语音信号的连续性注定了每帧语音与相邻语音的相关性，所以在静态特征的基础上获取语音帧间的时变特性可以提高识别的性能。

特征参数的Delta特征就是对该特征的语音帧序列的时序上作一次傅里叶变换，Delta特征也称为动态特征或2维特征。在具体实现时候会简化一些

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{3.1}$$

式中，d_t 就是第 t 帧的一阶 Delta 特征，Θ 表示计算 t 帧差分型特征所采用的语音帧的数量，一般选为 2～5 之间，在 HTK 中得默认值为 2，对于得到的结果再作一次相同变换就可以得到二阶参数，也就是 Delta-Delta 特征，也被称为速度系数（Velocity Coefficients），以此类推，可以得到三维特征，假若是 MFCC 参数，如果提取的是 16 维的 c_k，引入 Delta 特征，则最终的语音特征参数时 16*2=32 维：

$$x_k = \begin{pmatrix} c_k \\ \Delta c_k \end{pmatrix} \tag{3.2}$$

同理，再引入 Delta-Delta，则得到 16*3=48 维的特征参数。

使用TIMIT语音库里面的纯净语音，对其进行MFCC参数的提取，并提取出它的delta和delta-delta参数，三个参数的对比图如图3.1所示。

我们用HTK产生mfcc13（13维的MFCC）的mfcc26（13维的MFCC和13维的Delta

MFCC）和MFCC39（13维的MFCC、13维的MFCC Delta和13维的MFCC Delta-Delta）用对数字来进行语音识别得到的对比结果如表3.1所示

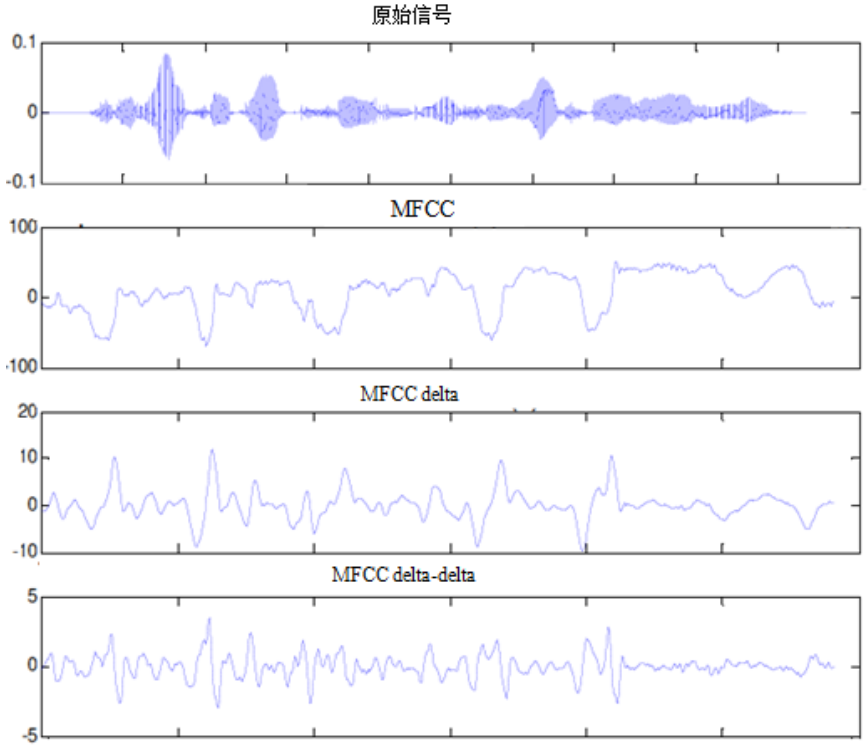


图3.1 语音信号的MFCC参数和MFCC的delta、delta-delta参数

表 3.1 13 维 MFCC 和它加上的 Delta、Delta-Delta 参数的识别准确率的比较

参数维数	干净语音		带噪声语音	
	Inside test	outside test	Inside test	outside test
mfcc13	82.59%	75%	79.24%	75.89%
mfcc26	91.29%	92.86%	83.71%	87.5%
mfcc39	91.07%	92.86%	84.6%	89.29%

根据实验数据得知，引入Delta特征能够在一定程度上提高识别率；Delta-Delta特征也可以提高系统的识别率，但是Delta-Delta比起Delta的识别率却并没有太大提高，有时候还会下降。

3.2 PCA

主分量分析(Principal Component Analysis, PCA)，又称为主元分析，是把多个指标化为少数几个综合指标的一种通用的统计分析方法对语音信号的特征经行降维，保留最有效的系数，舍去冗余的系数，在保证系统性能的同时，大大减少了计算量，提高了型号的实时性。对于某些受到噪音破坏的语音，其提取的特征因含噪音成分，不能精确的表征说话人，可以通过PCA变换，得到去除噪音成分的新的说话人特征。一般都是通过

KL变换实现的。PCA的变换流程为：计算出样本均值向量，对原向量作中心化变换，得到中心化后的向量序列，就中心化了原向量序列。再计算样本协防差矩阵。然后计算PCA转换矩阵，就可以得到新向量序列。

我们可以通过实验来看PCA的主轴方向是不是和数据是一致的，我们可以产生一个随机的数据，找出PCA的主轴，如图3.2所示

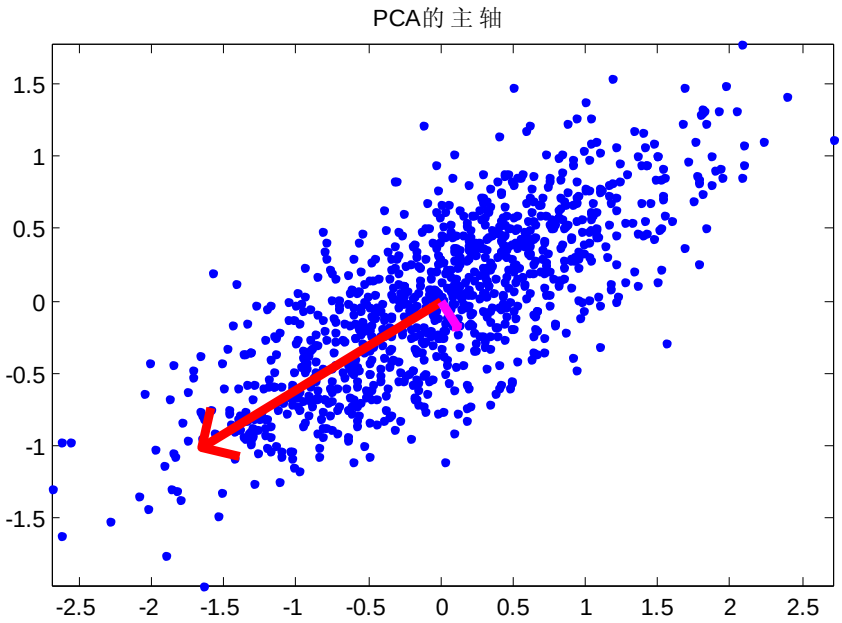


图3.2 PCA的主轴方向

很明显，PCA的主轴刚好的沿着数据分布最分散的方向。

基于kl的变换的PCA的实现步骤为：

（1） 平移坐标系，将模式总体的均值向量xm作为新坐标系的原点

$$x_t^* = x_t - x_m \tag{3.3}$$

其中

$$x_m = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{3.4}$$

（2） 求出总体自相关矩阵

$$R = \frac{1}{T}\sum_{t=1}^{T} x_t^* x_t^{*T} \tag{3.5}$$

（3） 求出R的特征值向量λ和对应的特征向量q。

（4） 将λ从大到小排序

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_M \geq \cdots \geq \lambda_N \tag{3.6}$$

取前M个大的λ所对应的q构成变换矩阵：

$$A=(q_1,q_2,\cdots,q_M) \quad M<N \tag{3.7}$$

（5） 将N维原始向量变换成M维新向量：

$$y_t=A^T x_t \tag{3.8}$$

其中，y_t的第一分类被称为原始向量x_t的第一主分量，包含了x_t中最多的信息，第二位分量称为第二主分量，依次类推。

3.2.1 随机向量的 k1 展开

X为N维的随机向量，则x可以用N个基向量的加权和来表示，x的kl展开式：

$$x=\sum_{i=1}^{N}\alpha_i\varphi_i \tag{3.9}$$

α_i是加权系数，φ_i为基向量，也可以用矩阵表示为：

$$x=\Phi\alpha \tag{3.10}$$

式中$\Phi=(\varphi_1,\ \varphi_2,\ \ldots,\ \varphi_N)$，$\alpha=(\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_N)^T$。

取基向量为正交向量

$$\varphi_i^T\varphi_j=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases} \tag{3.11}$$

则Φ为正交矩阵

$$\Phi^T\Phi=I \tag{3.12}$$

再考虑式（3.10）得

$$\alpha=\Phi^T x \tag{3.13}$$

寻找合适的Φ，使得α各个分量互不相关。

$$E[a_j a_k]=\begin{cases}\lambda_j & j=k\\ 0 & j\neq k\end{cases} \tag{3.14}$$

矩阵形式为

$$E[\alpha\alpha^T]=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \ldots & 0 & \lambda_N\end{bmatrix}=\Lambda \tag{3.15}$$

我们假设x的自相关矩阵为

$$R=E[xx^T] \tag{3.16}$$

则

$$R = E[\Phi\alpha\alpha^T\Phi^T] = \Phi E[\alpha\alpha^T]\Phi^T = \Phi\Lambda\Phi^T \tag{3.17}$$

两边同乘Φ，变换可得

$$R\Phi = \Phi\Lambda\Phi^T\Phi = \Phi\Lambda \tag{3.18}$$

也就是

$$R\varphi_j = \lambda_j\varphi_j \qquad j = 1,2,\cdots,N \tag{3.19}$$

式中，Λ_j是自相关矩阵的特征值，φ_j为对应的特征向量。

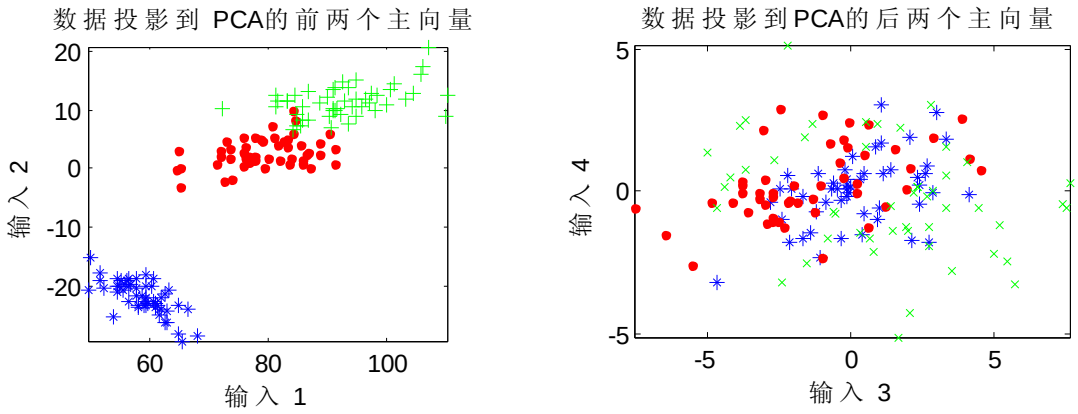


图3.3 数据在PCA上的投影

图3.3中第一个图是把数据投影到第一和第二个PCA主分量，第二个图是把数据投影到第三个和第四个PCA主分量。可以看出，在第一个图中数据分布比较分散，第二个图中数据的分散程度就比较小（第二个图比第一个图范围小）。

PCA的概念是将样本数据展开，并没有考虑到样本数据的类别，因此，严格来说，并不完全适合于样本识别的问题，但是由于“将样本数据展开”和“将不同类的样本数据展开”有一定的共同性，因此在数据维度太大的样本识别，如本文所介绍的说话人识别，PCA被用来降低维度和运算量。

3.2.2 k1 降维的实现

为了直线降维的目的，从N个特征向量中取出M个组成变换矩阵A

$$A = (\varphi_1,\varphi_2,\cdots,\varphi_M) \qquad M < N \tag{3.20}$$

这时A是一个N×M维矩阵，N维向量x经过变换可得M维的新向量。需要解决的问题是如何选取构成A的M个特征向量，才能使降维过程满足最小均方差准则。

取式(3.20)M项，对省略的项用预先选定的常数b_j代替，x的估计值表示为

$$\hat{x}=\sum_{j=1}^{M}\alpha_j\varphi_j+\sum_{j=M+1}^{N}b_j\varphi_j \tag{3.21}$$

产生的误差

$$\Delta x=x-\hat{x}=\sum_{j=M+1}^{N}(\alpha_j-b_j)\varphi_j \tag{3.22}$$

均方误差

$$\varepsilon^2=E[\|\Delta x\|^2]=\sum_{j=M+1}^{N}E[(\alpha_j-b_j)^2] \tag{3.23}$$

为了使均方误差最小，则选择应该满足

$$\frac{\partial\varepsilon^2}{\partial b_j}=0 \tag{3.24}$$

也就等同于

$$\frac{\partial}{\partial b_j}E[(\alpha_j-b_j)] \tag{3.25}$$

即

$$E[-2(\alpha_j-b_j)]=0 \tag{3.26}$$

由此可得

$$b_j=E[\alpha_j] \tag{3.27}$$

综合得到均方误差可变换为

$$\begin{aligned}\varepsilon^2&=\sum_{j=M+1}^{N}E[a_j{}^2]=\sum_{j=M+1}^{N}E[(\varphi_j{}^Tx)(\varphi_j{}^Tx)^T]\\&=\sum_{j=M+1}^{N}E[\varphi_j{}^Tx\varphi_j{}^Tx^T]=\sum_{j=M+1}^{N}\varphi_j{}^TR\varphi_j\\&=\sum_{j=M+1}^{N}\lambda_j\end{aligned} \tag{3.28}$$

从式中可以看出λ_j越小，ε就可以越小。通过对其从小到大的排序，然后取最大M个对应的λ_j特征向量组成变换矩阵，变换后得到降维的新变量，就能满足最小均方误差准则，是最优降维的结果。均方误差前提下的最优降维是PCA的基本属性之一，也是PCA能够得到广泛运用的原因。又因为新向量之间的互不相关，达到了去相关的目的，起到了减小相关性、突出差异性的效果。而且对于一些随机噪声，经过PCA转换后，会处于高频部分。适当的选取M，噪声就可以被除去。

我们实验中，在一个潜在的空间里用PCA找出数据的重要特征，并以此在原始数据空间里面来训练一个GMM模型，画出结果，如图3.4所示。

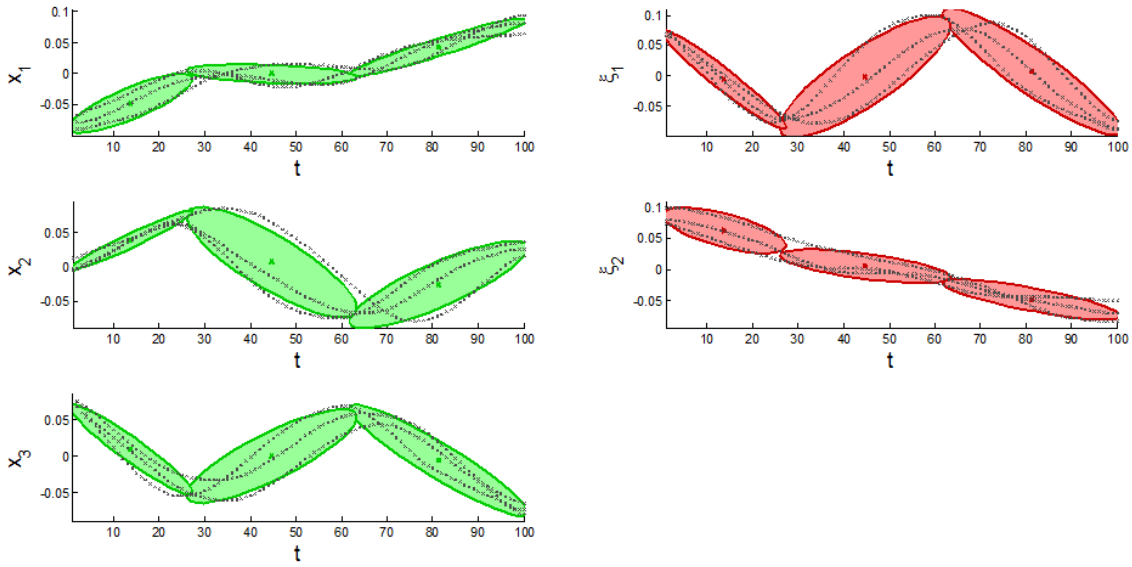


图3.4 原始数据与降维后数据GMM的比较

当数据是高维的时候用EM算法通常无法找到一个局部最优的解决办法，可以通过PCA把这些数据投影到一个潜在的空间的出低维参数作为一个预处理的步骤，再在原始数据的空间里面训练GMM模型。

PCA变换可以分为全局PCA变换和局部PCA变换，全局PCA就是所有说话人的训练特征作为一个总体，也就是整个系统只有一个PCA转换矩阵，每个人的特征都与这个唯一的转换矩阵相乘，得到新的特征向量。局部PCA是每个说话人都可以得到一个PCA转换矩阵。

3.3 谱减法与非线性谱减法

谱减技术是一种实用性很高的方法，它以噪声频谱是叠加在语音频谱之上的这一假设为基础，可以在语音的无声段采集噪声，因而可以先估计噪声频谱，而后从混噪语音频谱中减去噪声频谱。它是在假定语音和噪音是相互独立的基础上，首先估计出噪音能量谱,然后从含噪语音的能量谱中减去噪音的能量谱来得到干净语音的能量谱。我们假设

$$y(t)=s(t)+n(t) \tag{3.29}$$

式中，y(t)表示了含有噪声的语音；s(t)表示原始的干净语音；n(t)代表了加入干净语音里的噪音。

在频域里，由于事先假设语音跟噪音无关，因此含噪语音能量谱可以表示为：

$$S(w,t)=\hat{S}(w,t)+N(w,t) \quad 1\le t\le N \tag{3.30}$$

其中，其中S(w,t)是第t帧语音信号在W频带的能量谱，N(w,t)是t时刻w频带噪声能量的

估计，N 是FFT 的窗长。

我们用这种传统的方法来对一个语音信号进行去噪，语音采用的是TIMIT其中的一段语音片段，加的噪生是NOISEX-92数据库里的汽车噪声“Volvo”。对一段语音进行加强，这里认为第一帧的信号为噪声n(t)，运行结果如图3.5所示。

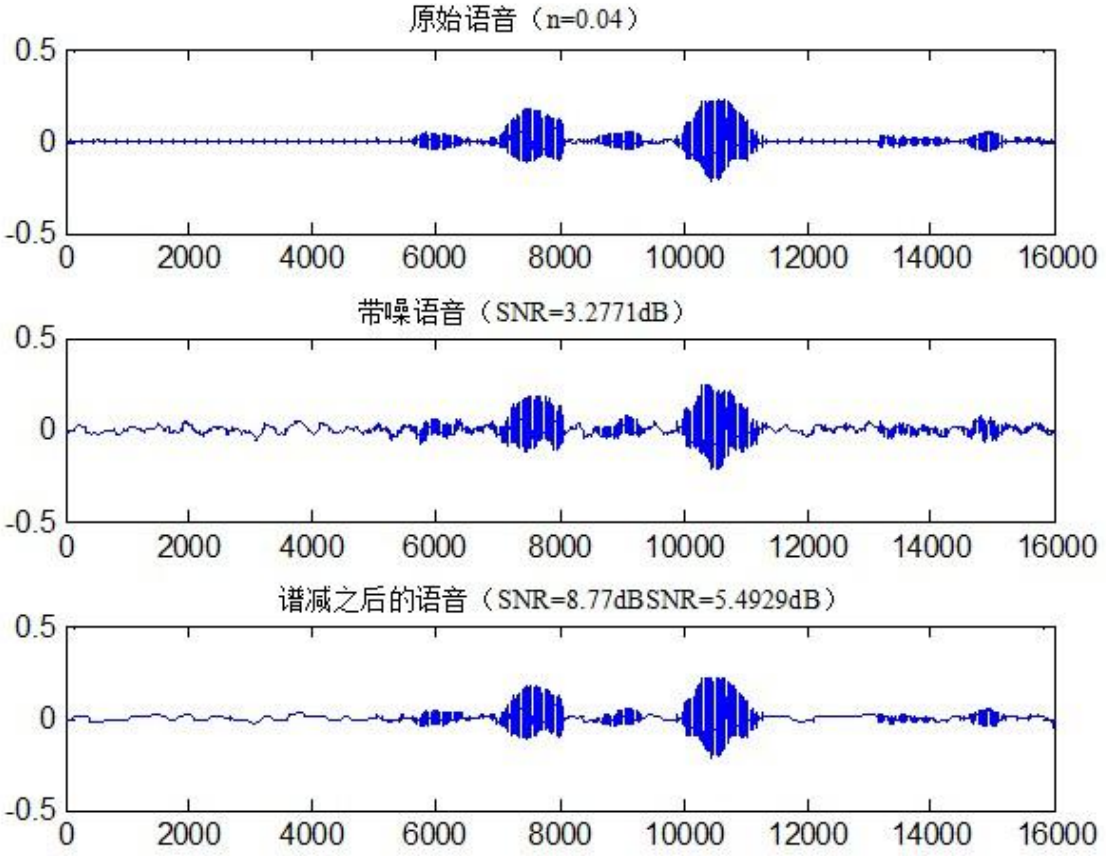


图 3.5 谱减法的语音增强

通常谱减会使得最终的能量谱出现负值，一般将其置为零值，然而从图中可以看出这样的处理会产生频谱尖刺，导致干净语音的能量谱产生尖峰和波谷，引入所谓的“音乐噪音”(Musical Noise)，影响识别效果。为了解决这一问题，Boll[54]在1979 年提出的谱减法中，将无语音段的能量谱平均值作为噪音能量谱的估计记为N(w,t)。干净语音能量谱可以通过下式得到

$$\hat{S}(w,t)=\begin{cases} S(w,t)-\alpha N(w,t) & S(w,t)>N(w,t) \\ \beta S(w,t) & \text{其他} \end{cases} \qquad (3.31)$$

式中，$\alpha \geq 1, 0<\beta<<1$

其中，α表示的是“过估计”(Over-Estimation)因子；ß表示的是“谱底”(Spectral Floor)因子。为了达到更好的效果，α可以定义为信噪比函数，信噪比高的部分需要较小的补

偿，信噪比低的部分需要较大的补偿。

3.3.1 噪声谱的估计

谱减法需要做预先的噪音估计，上式子的噪声能量谱一般的取得方法为：在对语音信号进行是否为语音段检测之后，就能估计出噪声的能量谱。当检测认定当前帧是非语音时，需要更新噪音能量谱。对所以的频带，噪声能力迭代更新公式为

$$N(w,t)=\begin{cases}\delta N(w,t-1)+(1-\delta)S(w,t) & QSNR<\theta \\ N(w,t-1) & \text{其他}\end{cases} \tag{3.32}$$

其中δ是控制噪声估计的更新速度，一般取决于当帧的信噪比；θ是判断语音的阈值，迭代的初始值定义为N(w,1)=S(w,1)，即假定第一帧是非语音。

由于噪音叠加的不可恢复性，谱减法会破坏原始语音的频谱。特别是当噪声较大时，谱减法往往会在抹去噪音的同时，也将语音的很多细节信息也抹去了，导致原始语音中有用信息的丢失，从而影响系统的识别性能。

3.4 倒谱均值相减法(CMS)

倒谱均值相减法(Cepstral Mean Subtract, CMS)[26]的基本思想就是：传输通道所带来的噪音环境一般是平稳，在倒谱域内卷积噪声的影响是加性的，所以将一段语音的倒谱参数减去这段语音的均值，就可以消除平稳卷积噪声的干扰。

$$X_d'[n]=X_d[n]-\frac{1}{N}\sum_{n=1}^{N}X_d[n] \qquad n=1,2,\cdots,N \qquad d=1,2,\cdots,D \tag{3.33}$$

式中，N 是该段语音特征的总的帧数，D是特征的维数。特征*Xd*[*n*]为第n帧的第d维特征分量。这种方法是消除倒谱域内噪音的一种简单且有效的方法，并且该方法对消除背景噪声也十分有效。通常倒谱均值相减法式作用整个语音文件上的。

3.5 倒谱均值与方差归一化法(CMVN)

语音信号经过加性噪音环境干扰之后，其倒谱的平均值与原本干净语音倒谱平均值之间通常存在这一个偏移量。同时，由于噪音的存在，必然使得语音倒谱参数的方差相对于干净语音参数的方差缩小。这样如果训练和测试环境不一样，那么就会使得特征的不匹配，从而降低了识别效果。倒谱均值与方差归一化(Cepstral Mean and Variance Normalization, CMVN)[27]的思想是把每一维的倒谱特征参数的平均值归一化为0，并且将其方差归一化为1，这样的话就可以降低上述的不匹配，从而增强特征参数的鲁棒性。其具体方法为：先使用倒谱均值相减法(CMS)做处理，每一维倒谱系数的平均值为0。

然后再对处理之后的每一维倒谱系数除以其标准差，这样就可以得到新的特征参数。其公式如下：

$$X'[n]=\frac{X[N]-u_x}{\delta_x} \qquad n=1,2,\cdots,N \tag{3.34}$$

其中，$u_x=\frac{1}{N}\sum_{n=1}^{N}X[n]$为倒谱均值，$\delta_x=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(X[n]-u_x)^2}$ 为标准差。

3.6 RASTA

RASTA(Relative Spectral Processing)是由H.Hermansky[28]等提出的一种鲁棒处理方法。我们知道信道和噪音等等会影响语音信号的短时频谱，会使之产生偏移。但是这些信息往往是固定的或者是缓慢变化的，他们的变化率与普通语音信号的变化率范围有很大差别。感知实验指出，对于缓慢变化的声源信息人类的听觉感知是非常不敏感的，因此人类的听觉本身可以说是抑制缓慢变化的非语音信息的，同时增强变化较大的语音信息。RASTA的思想就是利用人耳的这听觉特性，对语音参数时间域内进行滤波，去掉较稳定的通道或噪音信号。

RASTA滤波的处理方法是用一个带通滤波器对语音参数进行滤波，滤掉低频的通道噪声以及高频的噪声干扰等。处理后频谱中的缓慢变化的信息被抑制，人耳能够感的动态语音被保留下来。滤波器的频响公式如下：

$$H(z)=0.1\times\frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}\times(1-0.98z^{-1})} \tag{3.35}$$

相当于一个带通滤波器，它的高低截止频率决定了所保留和去除的谱的变化。RASTA滤波器和语音特征参数(RASTA-PLP[28],RASTA-MFCC[29]等)的结合能够得到更良好的识别效果和鲁棒性。

3.6.1 特征提取的 RASTA-PLP 技术

特征提取的整个过程如图3.6所示[16]。PLP特征分析是基于短时谱的，并且是基于人耳听觉的特征，所以PLP和其他基于短时谱的参数都会在进行分帧时引入帧移谱变化，使用RASTA方法可以抑制这种变化使得PLP和其他的一些基于短时谱的技术对线形谱失真更具有稳健性。

对PLP特征提取与RASTA滤波相结合的过程是：在过程图中可以看到，PLP特征提取与RASTA滤波相结合的过程就是在原PLP特征提取的过程中再加了三个步骤，首先对语音经行临界频带分析，在对临界带功率谱幅度进行非线性压缩变换（取对数）之后使

用RASTA滤波，然后再对数据进行非线性的扩展变换（取反对数），之后再按照原来PLP特征提取的步骤继续进行。

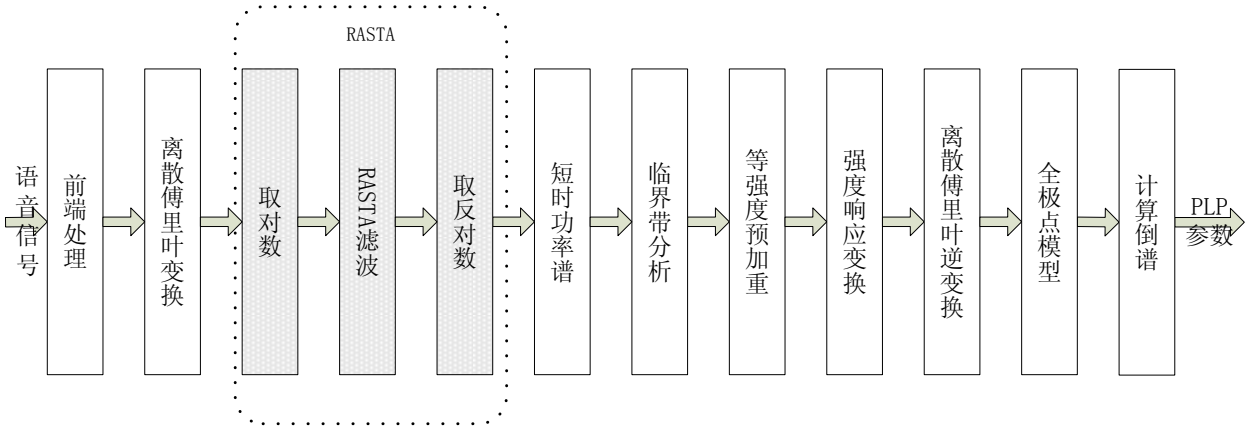


图3.6 语音信号RASTA-PLP参数提取过程

非线性压缩函数取自然对数In x，扩展函数取指数e^x。

我们仍然使用TIMIT的语音和NOSIE-92的汽车噪声，对原始的纯净语音进行PLP参数的提取，画出其谱图，然后再对带噪的语音分别进行PLP参数提取和经过RASTA滤波的参数提取，画出谱图，与纯净语音的谱图相比较，结果如图3.7所示。

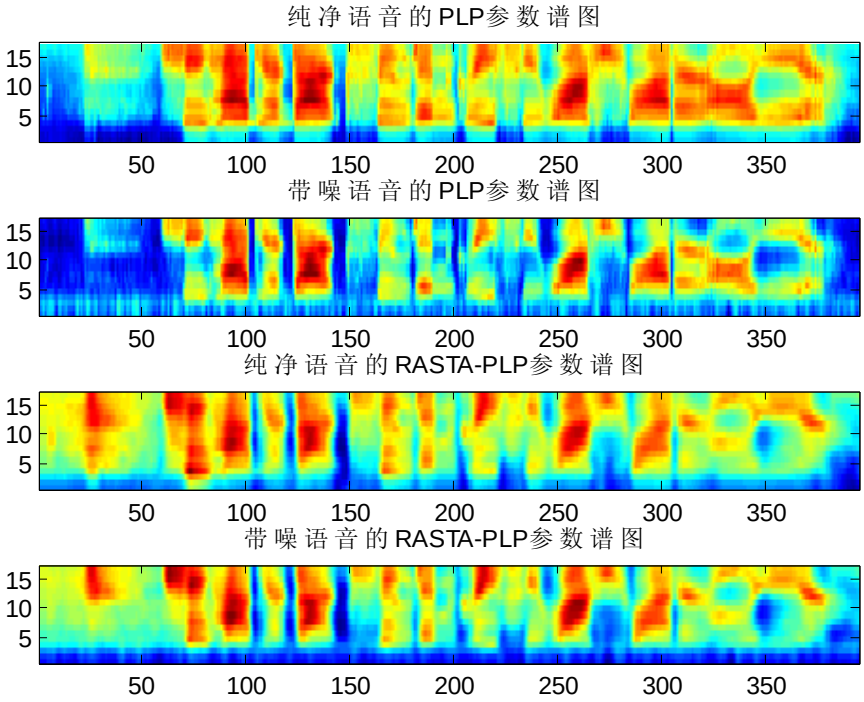


图3.7 纯净语音的PLP，RASTA-PLP和
带噪声语音的PLP，RASTA-PLP参数的比较

3.7 特征弯折(Feature Warping)

训练和测试的过程中，由于通道和噪音的影响会造成训练和测试语音的参数在概率分布上差别较大，导致不够匹配。在说话人确认中，我们认为同一个说话人所提取的特征值应该有固定的范围，从物理上看，语音帧的同维倒谱参数的分布应该是正态分布的。特征弯折(Feature Warping)[30]是在说话人识别中提取出来的特征归一化方法，其的目的就是使得训练和测试语音的特征寻列通过积分布函数(Cumulative Distribution Function, CDF)变化为符合标准正态分布的特征序列，都映射到相同的统计分布空间，从而提高对不同通道和噪音的鲁棒性。特征弯折是在假设倒谱特征各维独立的情况下计算的，因此可以分别对各维经行单独的处理。这原始特征从左至右已排好序，弯折就是将其原始参数映射到一个固定的范围内，并且在这个范围内的分布曲线是标准正态分布。这样训练语音和测试语音的概率分布都逼近了同一个参考概率分布，降低噪声对原始语音影响导致的变异，提高了鲁棒性。

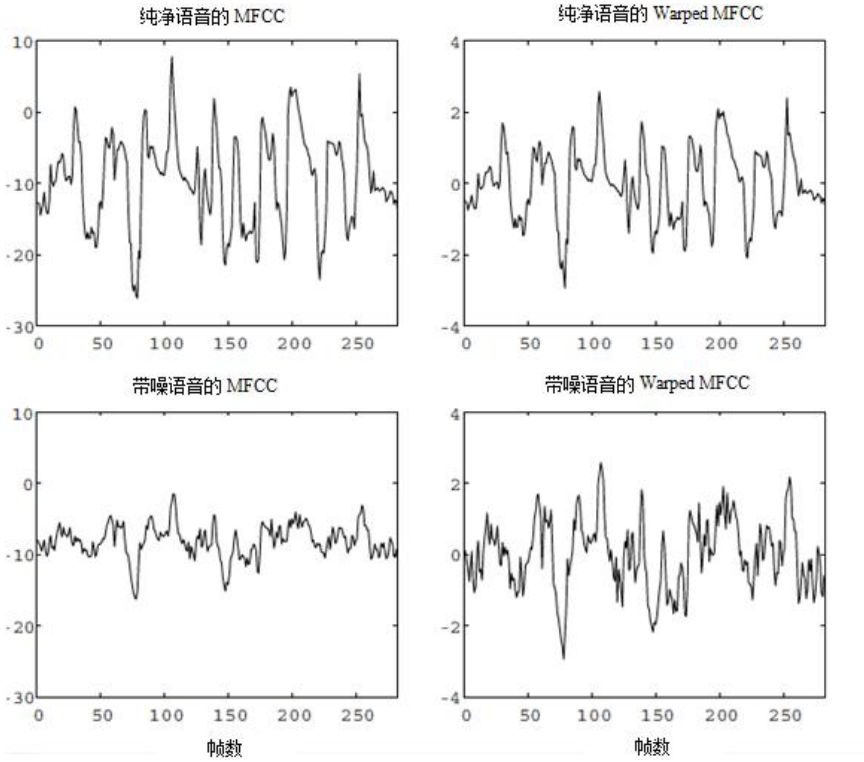


图3.8 纯净语音的MFCC、特征歪曲的MFCC
和带噪语音的MFCC、特征歪曲的MFCC之间的比较

具体处理方法为：首先对窗长为N里的每段语音视为一个滑动窗口(sliding window)分别进行特征弯折。对这个窗内的同维倒谱参数值进行从大到小的降序排列。如果原来处于滑动窗某位置的倒谱系数值排序后的位置为R(在1和N之间)，按照公式进行处理就可以得到特征弯折后对应的倒谱参数值m。

$$\frac{N+\frac{1}{2}-R}{N}=\int_{z=-\infty}^{m}h(z)dz \qquad (3.36)$$

其中h(z)为标准正态分布曲线

$$h(z)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2}) \qquad (3.37)$$

用纯净语音和带噪语音(SNR=5)的原来的MFCC参数和使用特征弯曲的MFCC参数进行高斯目标映射，得到结果如图3.8所示。

3.8 本章小结

介绍了参数级的鲁棒性算法：动态参数法、既可以降维又可以去噪的PCA、谱减法（SS）、倒谱均值相减法(CMS)、倒谱均值与方差归一化法(CMVN)、RASTA 和特征弯折(Feature Warping)等，这些参数级算法被广泛采用。与特征弯折类似的还有短时高斯化和异方差归一化等算法，这种算法和谱减、倒谱方差归一的方法都不依赖噪声的先验信息，实现要更加方便，但是不能完全很好的消除噪声使语音信号产生的变化。也有很多文献提出了本身就具有抗噪特性的特征参数[18、25]，大多都是多几种特征参数和参数处理方法的融合[24、46]，还有的融入了小波[23、50]，都取得了不错的效果。本文在第四章也提出了一个基于特征参数融合的抗噪的特征参数的提取算法。

4 基于 GMM-UBM 说话人识别模型

4.1 GMM 模型

从上世纪九十年代以来，高斯混合模型(Gaussian Mixture Model, GMM)[4]作为一种通用的概率模型被广泛应用于与文本无关的说话人确认中，它能有效地模拟多维矢量的任意连续概率分布，如图4.1所示能很好地刻画参数空间中训练数据的空间分布及其特征，因而很适合文本无关的说话人识别，并且具有简单高效的特点。

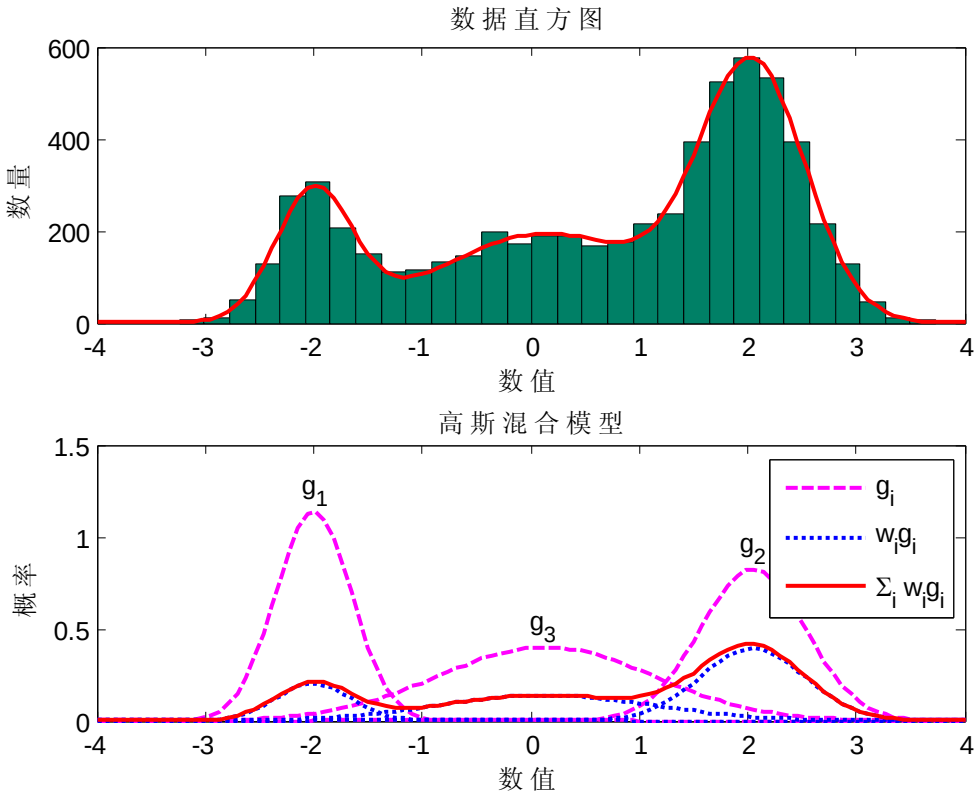


图4.1 GMM对数据分布进行逼近

声纹识别的统计概率模型方法主要有隐马尔可夫模型(Hidden Markov Model, HMM)和高斯混合模型。HMM 和GMM 都可以用于与文本无关的声纹识别系统。HMM 模型参数不仅包含了训练语音的声学特征还包含了训练语音随时间变化的状态信息，但HMM参数估计复杂而且系统训练需要大量的语音数据，不利于实时识别。研究表明，在与文本无关的识别中，GMM的识别性能要比HMM 好。

GMM 可以被视为是一个状态数为一的连续分布的隐马尔可夫模型CDHMM

(Continuous Density Hidden Markov Model)，通过若干个高斯概率密度函数的线性组合来逼近任意分布，从而模拟出各种形式的语音特征分布，描述出说话人个性的特征的统计分布，以区分不同的说话人。GMM 有很高的高斯混合度，GMM 中，它是通过来从说话人语音抽取出来的特征矢量X对应的似然率可以用M个高斯混合概率密度分布的加权组合表示[47]：

$$P(X/\lambda)=\sum_{i=1}^{M}w_i b_i(X) \qquad i=1,2,\cdots,M \tag{4.1}$$

在式中，X是阶数为D维的特征矢量，$b_i(X_t)$代表了子分部，M为是高斯混合分布的阶数，也就是单高斯分布的个数，高斯混合的个数与语音识别的关系如图 所示；w_i是第i个高斯分布的权重。用HTK工具箱来算出在不同上面提到的mfcc13、mfcc26和mfcc39的特征参数情况下GMM的个数（从1到18）和识别率之间的关系，如图4.2所示

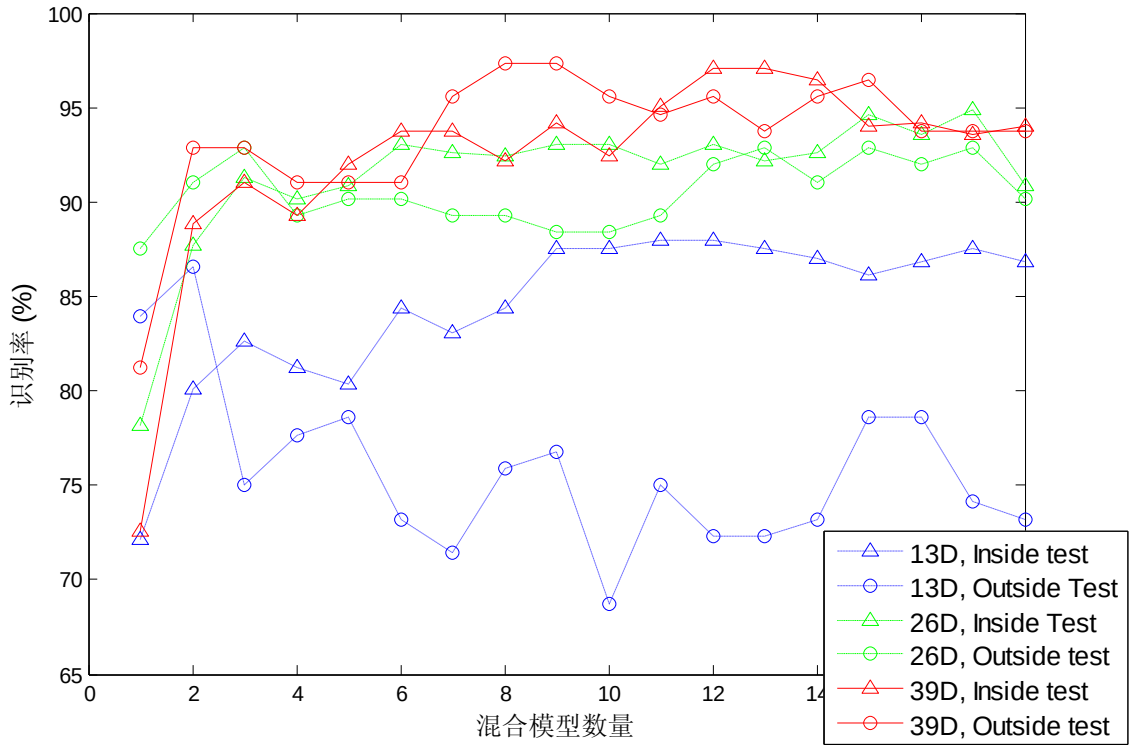


图4.2 GMM的个数和识别率之间的关系

D维的$b_i(X)$可表示为

$$b_i(X)=N(w_i,\mu_i,\Sigma_i)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(X-\mu_i)^t\Sigma_i^{-1}(X-\mu_i)\right\} \tag{4.2}$$

式中，μ_i是均值向量，Σ_i是协防差矩阵，混合权重w_i满足

$$\sum_{i=1}^{M}w_i=1 \tag{4.3}$$

协方差矩阵Σ_i可以用完全的协方差矩阵，从理论上也认为它可以更加合理的描述说话人特征分布。但是通常情况下，语音的特征参数维数一般较大，如果用全矩阵协方差将导致模型参数数目过多，一方面由于数目过多，在有限的训练语音下不可能可靠地估计，而且训练过程中需要用到协方差矩阵的逆,这样对计算量要求太高，一个M阶的完全协方差高斯混合模型可以用更高阶的对角协方差高斯混合模型代价表示，如果采用对角协方差，计算量可以变小。因此，在实际使用中出于参数数量与计算量的考虑，通常是假设特征矢量各维不相关，采用对角形式的协方差。也就是认为非对角元素为零。而且有研究指出采用对角协方差的比采完全的协方差的模型的识别性能要好。

一个完整的GMM模型可以表示为

$$\lambda=\{w_i,\mu_i,\Sigma_i\}\quad i=1,2,\cdots,M \tag{4.4}$$

GMM模型求得的对数似然度定义为

$$L(X/\lambda)=\frac{1}{T}\sum_{t=1}^{T}\log P(X_t/\lambda) \tag{4.5}$$

GMM 的基本结构以及观察特征矢量与模型匹配的基本示意图如图4.3所示

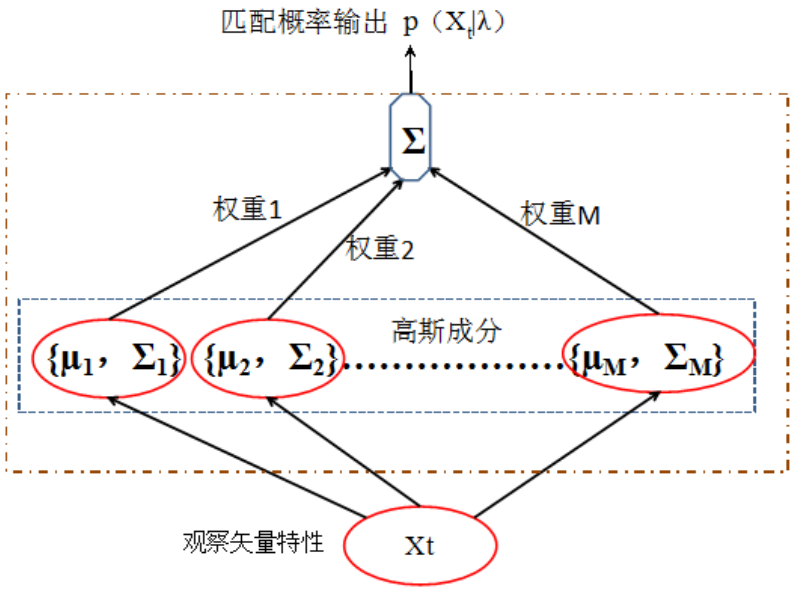


图4.3 GMM的基本结构以及观察特征矢量与模型匹配的基本示意图

4.1.1 GMM 的参数估计

在GMM模型确定参数时最常用的就是最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)估计，可表示为

$$P(X/\lambda)=\prod_{t=1}^{T}P(X_t/\lambda) \tag{4.6}$$

MLLR模型的参数可以用EM(Expectation Maximization)算法估计。EM算法的初始值的确定有不同的方法，确定的值不同。主要有三种方法：随机初始化(Random Initialisation)、K-means初始化(K-means Initialisation)、迭代初始化(Iterative Initialisation)。为了对比这三种初始算法的效果，我们使用TIMIT语音库里面的20秒长度的语音来对一个GMM进行训练，每种情况训练25次。图4.4描述出了三种情况在进行不同次数EM（最高的是21次）的迭代之后规范化的负对数似然度的平均值(normalised negative log likelihood)，很明显可以看到不同的方法所找出的局部最大值并不一样。

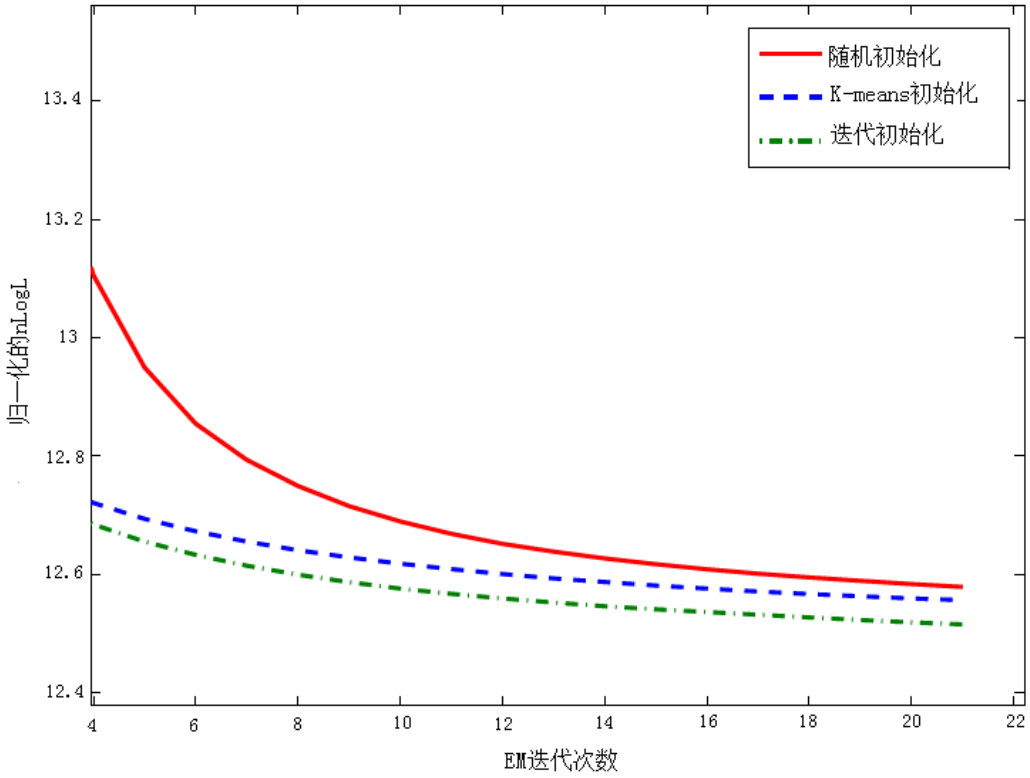


图4.4 三种初始算法的收敛情况

为了提高高斯模型输出训练矢量的似然分，用EM算法迭代重估高斯混合模型的参数，下面列出GMM模型各个参数重估公式。

混合权重：

$$\omega_i = \frac{1}{T}\sum_{t=1}^{T} P(i/X,\lambda) \tag{4.7}$$

均值：

$$\mu_i = \frac{\sum_{t=1}^{T} P(i/X_t,\lambda)X_t}{\sum_{t=1}^{T} P(i/X_t,\lambda)} \tag{4.8}$$

方差：

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} P(i/X,\lambda)(X_t - \mu_i)^2}{\sum_{t=1}^{T} P(i/X,\lambda)} \tag{4.9}$$

后验概率为

$$P(i/X,\lambda) = \frac{w_i b_i(X_t)}{\sum_{k=1}^{M} w_k b_k(X_t)} \tag{4.10}$$

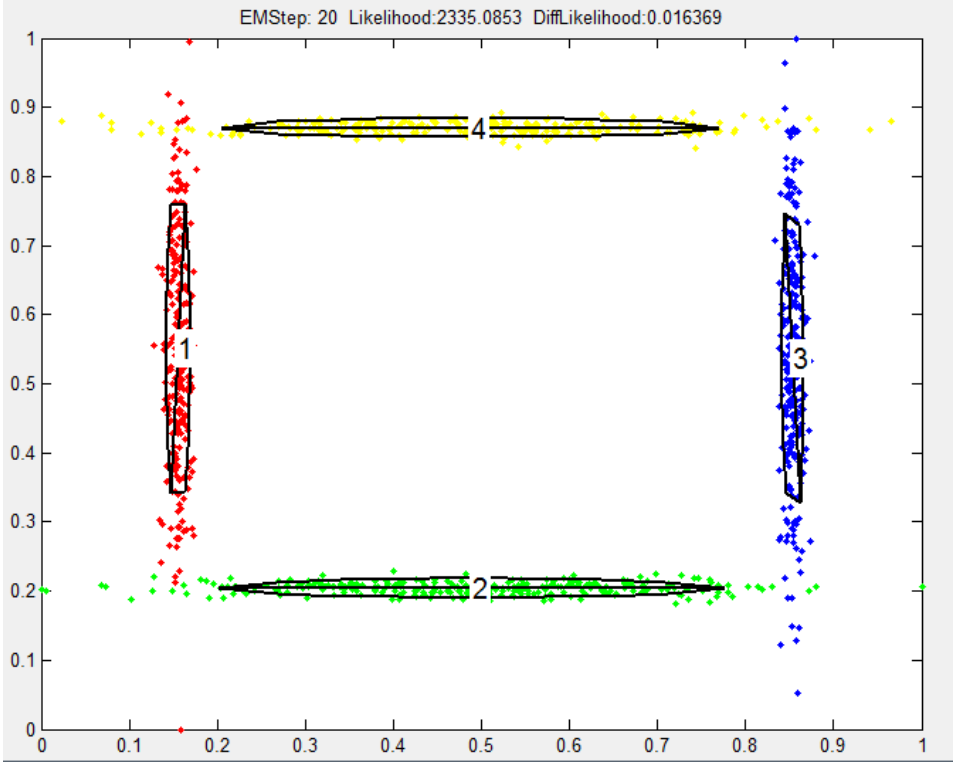


图4.5 GMM对数据进行分类

GMM既具有参数模型的有结构和参数控制函数密度的行为，但却又有非参数模型的自由度，很适合用于对语音这样的时变信号经行建模。与文本无关的说话人识别也就代表了用于训练和测试的语音的语义完全可以不同。而在实际情况中，对于某个特定的说话人，由于文本不同、intra-speaker的生理变化及流体动力变化差异等等因素的影响，他的发音器官如声道形状和声门激励波都不可能一样。也就是语音的产生过程是不确定的。这样，我们提取的说话人特征矢量的分布并不严格服从某一种解析概率密度函数。但是，由概率论的中心极限定理可知，大量相互独立、服从同一分布的随机变量在总体上服从正态分布。并且，任何一种概率密度分布均能由若干个高斯密度函数(正态密度分布)的线性加权组合来逼近。由于正态分布能够反映自然界中大多数事物的分布规律，因此得到了最广泛的应用。具体地说，我们可以通过每一个高斯分布来描述每一个单音

素(或者更细的划分)以及它的变化范围。理论上说无限多个高斯分布就可以拟合一个任意形式的分布，所以当GMM 的混合度M足够多时，可以描述足够多的音素，从而就可以足够精细地描述说话人特征参数的空间分布。而GMM 模型刻画了每个人语音特征参数的静态分布，不同说话人的语音特征的静态分布也是不同的。因此通过比较不同说话人的GMM 模型，就可以区分不同的说话人。

4.2 一种新的抗噪特征参数用于 GMM 模型的说话人识别

在上一章介绍了很多的特征参数的补偿方法，这里提出了一种具有噪声鲁棒性的特征参数提取的方法，这种算法的基本思想就是利用 LPCC 可反映的共振峰特性、MFCC 符合人耳听觉特性和自相关函数的较好抗噪性能，参数的优点相结合，得出一个新的抗噪语音特征参数。并将这种算法与 GMM 模型相结合，对说话人进行确认。

4.2.1 参数的提取

（1） 短时自相关函数

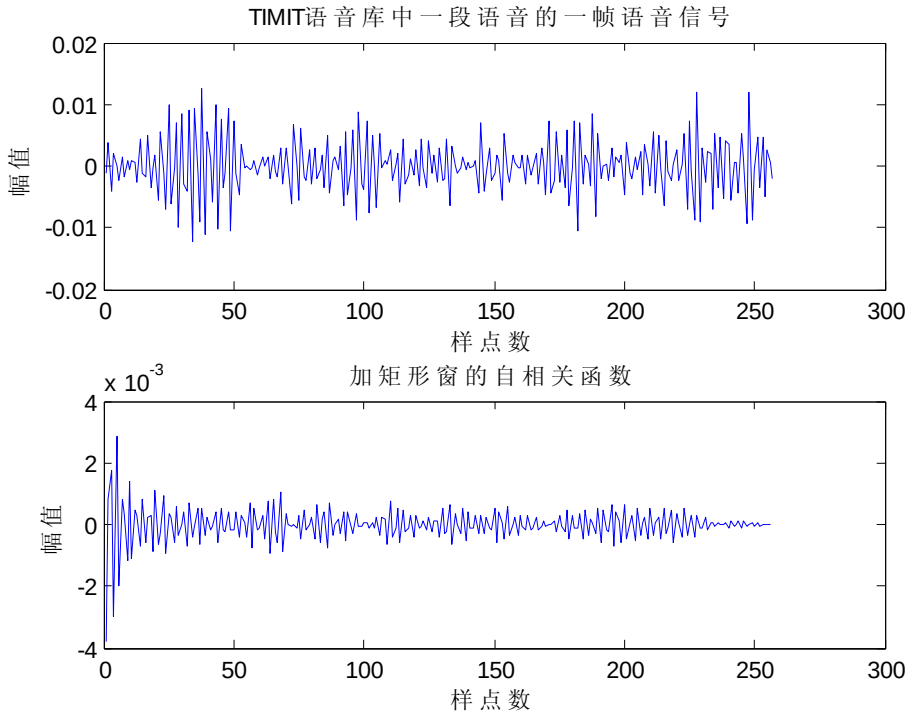


图 4.6 一帧语音信号的短时自相关函数

假设一段语音信号 x(m)，语音信号短时自相关函数的定义为：

$$R_n(k)=\sum_{m=0}^{N-1-k}[x(n+m)w'(m)][x(n+m+k)w'(k+m)] \quad (4.11)$$

其中的w'(m)为加窗函数，n为帧移的长度。使用TIMIT语音库中的一段语音，以矩形窗为例，一般情况下语音信号的窗长取20ms至40ms，以本文所用的语音为例子的话，本文所用语音采样频率是1600Hz，如果取窗长为24ms的话，那么窗长N就为256，帧移就可以选取为128。图4.6所示为语音中一帧的信号和它的自相关函数。

（2） 对自相关函数进行线性预测

线性预测的也就是认为 Rn(k)的值是前 m-1 个数值的线性组合得到的，又因为 Rn(k)是由气流 H(n)冲激声门发出的语音信号 S(n)计算得到，所以我们得到以下的公式：

$$R_n(k)=G^*H(k)+\sum_{l=1}^{m}a_l R_n(k-l) \quad (4.12)$$

两边同时 Z 变换

$$R_n(Z)=G^*H(Z)+\sum_{l=1}^{k}a_l R_n(Z)Z^{-1} \quad (4.13)$$

$$H(Z)=\frac{R_n(Z)}{X(Z)}=\frac{G}{1-\sum_{l=1}^{k}a_l Z^{-l}} \quad (4.14)$$

根据 z 变换

$$H(Z)=\sum_{n=+\infty}^{n=+\infty}h(n)z^{-n}=\sum_{n=0}^{n=+\infty}h(n)z^{-n} \quad (4.15)$$

由上面两式可得

$$h(n)=\begin{cases}a_n+\sum_{l=1}^{n-l}(1-l/n)\omega h(n-1) & (1\le n\le k)\\ \sum_{l=1}^{n-l}(1-l/n)\omega h(n-1) & (n>k)\end{cases} \quad (4.16)$$

式中，h(n)就是线性预测系数，ω 通过最小二乘法得到。

（3） 线性 Mel 尺度变换

$$MC_k(n)=\begin{cases}h(n)+\sum_{p=1}^{m}\alpha_p MC_0(n+p) & k=0\\ \sum_{p=1}^{m}\alpha_p MC_{k-1}(n+p)+\sum_{p=1}^{m}\beta_p MC_k(n+p) & k>0\end{cases} \quad (4.17)$$

式中

$$\sum_{q-1}^{m}(\alpha_p+\beta_p)=1 \tag{4.18}$$

$$\alpha_p=\frac{\alpha_p'}{\sum_p \alpha_p'} \tag{4.19}$$

$$\beta_p=\frac{\beta_p'}{\sum_p \beta_p'} \tag{4.20}$$

$$\alpha_p'=\frac{1}{1+\exp(p)} \tag{4.21}$$

$$\beta_p'=1-\alpha_p' \tag{4.22}$$

MCk 为 Mel 系数，n 为迭代的次数，K 为 Mel 系数的阶数，p 为尺度变换的阶数。

4.2.2 实验结果

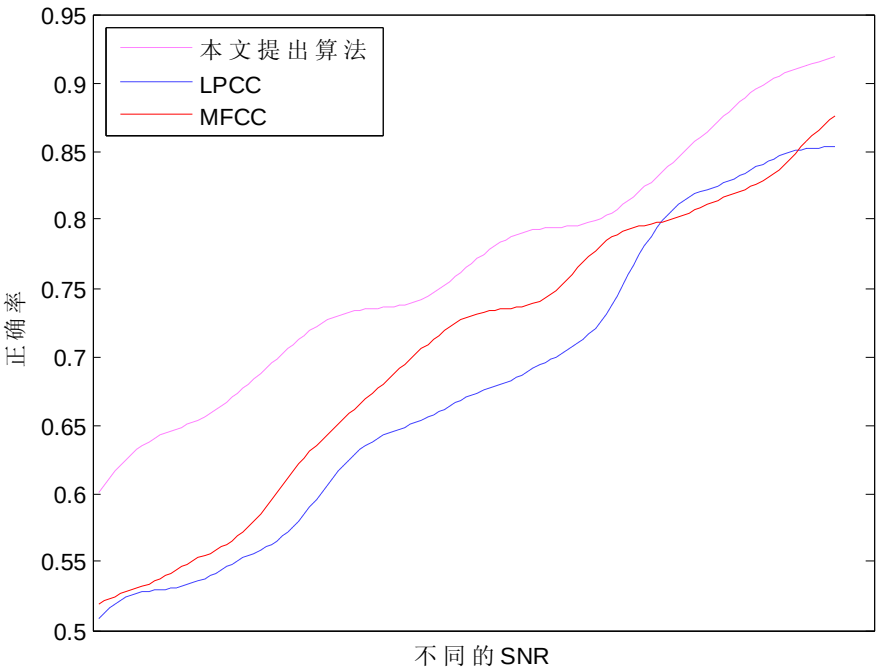


图 4.7 三种参数用于 GMM 模型的说话人识别的比较

把得到的参数用于训练 GMM 模型来识别语音，用 TIMIT 的数据库中 10s 的语音信息再加入 NOISE-92 中的噪声来训练识别，将提出的参数提取方法与 MFCC，LPCC 相比较。

在不同信噪比下比较系统识别的准确率，识别结果如图 4.7 所示。从图中可以看出，在纯净语音中加入不同信噪比的噪声，系统所得的正确率要明显高于 LPCC 和 MFCC，证明了这种算法的有效性。

4.3 基于高斯混合模型-通用背景模型（GMM-UBM）的说话人确认

在NIST主办的说话人确认评比中，性能在前几名的系统基本都是基于GMM-UBM的框架结构。在GMM的说话人识别中，说话人模型通常用说话人的很短时间的语音训练得到，训练参数不是很充足，但是这个有限的训练语音并不能覆盖特定说话人所有可能的发音情况。当测试语音不同于训练语音的时候，就会因为不能很好地与模型匹配从而会影响识别率。UBM 是一个说话人无关、高阶的高斯混合模型。该模型通常由很多不同人的大量语音训练得到，用于表示说话人的统计平均发音特性。基于UBM建立的GMM模型就可以很好的弥补GMM的不足，提高系统的识别率。

基于GMM-UBM 的系统有两个好处：

（1） 说话人模型是在UBM 上根据说话人的训练语音自适应得到的。这样，对于说话人训练语音覆盖到的发音，可以用该说话人自身的语音建模；对于未覆盖到的发音，可以用UBM 里的发音分布近似，从而减少测试语音与训练语音在声学空间上由于分布不同所带来的影响。

（2） UBM 可以被看作是一个“标准参考者”的模型，这样在进行身份确认的时候，可以用测试语音在UBM 上的得分来作为一种参考阈值。

基于GMM-UBM的说话人主要分为：UBM训练，说话人模型自适应，说话人确认测试[14]。本文分为训练过程和识别过程两部分来介绍。

4.3.1 基于 UBM-GMM 说话人确认的的训练过程

我们知道，在概率统计的GMM系统中，当混合度M足够多时，可以足够精细地描述说话人特征参数的空间分布。一般情况下，如果GMM的混合度越高，系统的识别性能就越好。所以我们期望能够尽可能地训练高混合度的GMM。但是训练高混合度的GMM 需要足够多的语音数据，需要大量的训练数据集，这样才能够反应说话人的特征参数的空间分布。D.A.Reynolds 就提出了基于UBM-GMM的与文本无关的说话人识别系统[15]。统一背景模型(Universal Background Model, UBM)是一个混合度非常高的

GMM。它是采用了几百人、涉及不同信道的大量语音训练一个高阶的GMM，描述了说话人无关的、包含多种信道的特征分布空间。UBM 并不是描述某个特定人的特征分布，而是描述了包含有各种通道和噪音的很多人的语音特征空间，所以称之为背景。

在实际的使用中，在有了大量的给定数据之后，就可以训练UBM模型。如图4.8所示。

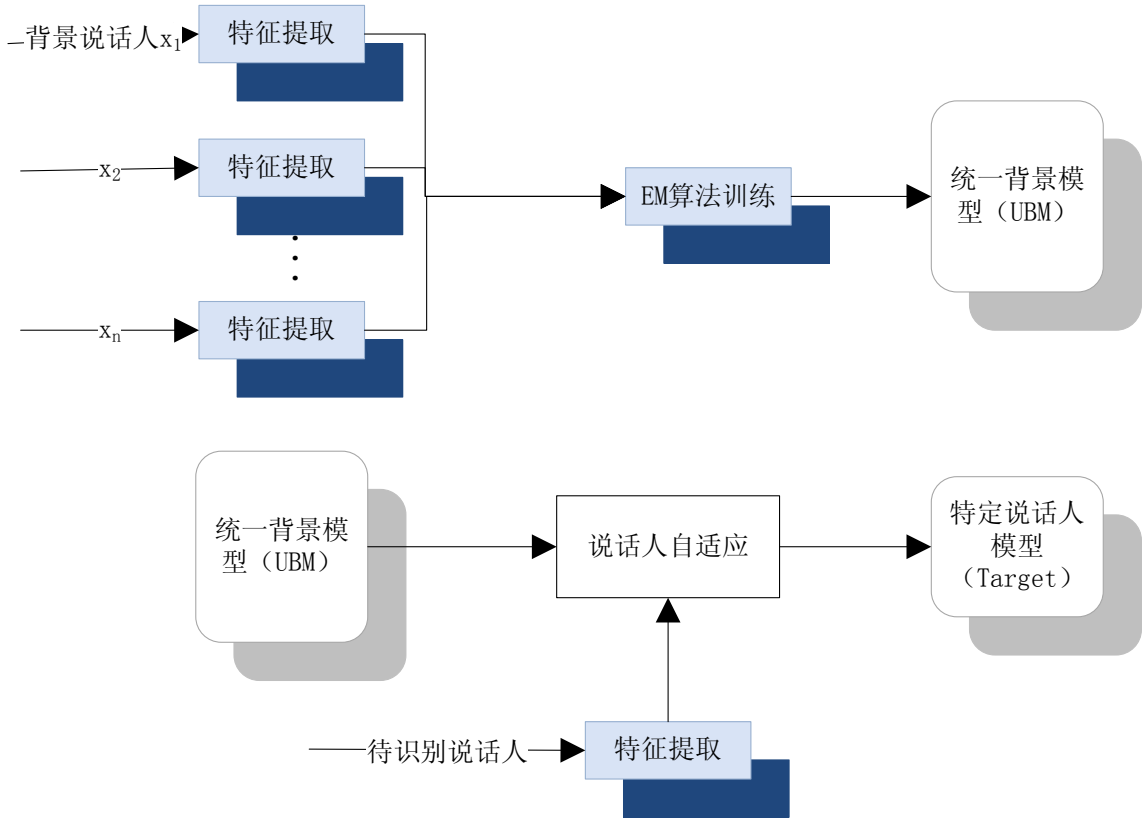


图4.8　UBM训练和说话人模型自适应过程

一般使用的也就是最简单的方法就是把所以数据放到一起采用EM算法来估计。使用这种方法时需要注意保证训练数据之间的比列平衡，如男，女语音的数量还有年龄比列等。考虑到男女之间的语音差异较大，一般UBM 是分性别(Gender dependent)而进行训练的，通常对于男女分别训练一个UBM。对于该UBM，使用最大后验概率自适应算法得到特定说话人的模型(Target Model)。在特定说话人的模型的训练中，GMM-UBM的说话人模型是通过自适应方法修改了以UBM 模型作为初始模型某些参数而得到的，一般来说，常用的自适应算法有贝叶斯最大后验概率估计(Maximum A Posteriori, MAP)算法和最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)算法。前者需要估计出某一特定环境下的先验模型参数，从而对说话人模型进行相应的补偿；后者假定用一小部分语音数据即可估计出训练环境与测试环境之间在模型参数上的差异，在此基础上，对说话人模型进行修正。在论文使用的实验系统中，说话人模型自适应所使用的算法是MAP。对GMM-UBM模型的训练过程分为两步，首先是期望过程，也就是按照EM 算法中重估参数的公式统计出训练数据对UBM各单高斯分布上计算出特定说话

人的最优统计模型参数的估计；第二步就是在得到训练数据的最优参数之后，用一个语料相关的混合系数将新的统计参数加权之后对初始模型UBM的参数进行修正，加权是为了使最终的说话模型的参数分布比起UBM的参数要更接近说话人自身的参数。完成这一步时，因为特定说话人的语音训练样本有限，不同于EM 算法中直接用重估好的参数代替初始模型的参数的方法，基于MAP自适应算法中，为了使重估的数据更加可靠，不能只靠训练数据，需要得出估计训练数据落在某个混合度上的个数再进行判决。

假设特定说话人的训练语料为x_t，它特征向量对高斯分布的后验概率值为：

$$p(i \mid x_t) = \frac{w_i b_i(x_t)}{\sum_{k=1}^{M} w_k b_k(x_t)} \tag{4.23}$$

加权值：

$$n_i = \sum^{T} p(i \mid x_t) \tag{4.24}$$

平均值向量：

$$E_i[x] = \frac{1}{n_i} \sum_{i=1}^{T} p(i \mid x_t) x_t \tag{4.25}$$

平均值方差：

$$E_i[x^2] = \frac{1}{n} \sum_{i=1}^{T} p(i \mid x_t) x_t^{\,2} \tag{4.26}$$

用这些统计量来对 UBM 里的参数进行更新

$$\hat{w}_i = [\alpha_i^{\,w} \frac{n_i}{T} + (1-\alpha_i^{\,w}) w_i]\eta \tag{4.27}$$

$$\hat{\mu}_i = \alpha_i^{\,m} E_i[x] + (1-\alpha_i^{\,m})\mu_i \tag{4.28}$$

$$\hat{\sigma}_i^{\,2} = \alpha_i^{\,v} E_i[x^2] + (1-\alpha_i^{\,v})(\Sigma_i^{\,2} + \mu_i^{\,2}) - \mu_i^{\,2} \tag{4.29}$$

式中自适应参数 α_i 被用于控制新旧参数的平衡；γ 为权重的规整因子，使得自适应之后素养的的权重之和为 1；用于上面的自适应参数 α_i^{w}，α_i^{m}，α_i^{v} 分别为权重，均值，方差的调整系数，定义为：

$$\alpha_i^{\,\rho} = \frac{n_i}{n_i + r^{\rho}} \qquad \rho = w, m, v \tag{4.30}$$

式中 r^{ρ} 为 MAP 算法中的先验设置，是一个固定值（一般为 16），控制 UBM 的参数在自适应中的权重，采用数据相关的自适应参数是为了保证自适应是和高斯分布的相关性；n_i 表示了落在第 i 个高斯混合度上的训练矢量的个数,n_i 的引入保证了系统的可靠性，我们可以从式中看出随着 n_i 变大 α_i^{ρ} 就越趋近一，第 i 个高斯混合度重估的参数越可靠，这个参数能对模型的参数的更改的权重也越大。

特定说话人的GMM-UBM模型是在UBM上根据说话人的训练语音自适应得到的，是一个与UBM之建立了对应的关系且与UBM 混合度相同的GMM。在训练时，用特定说话人的特征参数修正UBM中描述了非特定人的语音分布的参数。在这个过程中，保留了UBM中描述通道以及背景噪声的参数，使在训练数据有限的情况下建立的模型具有鲁棒性。而且由于在训练修正过程中使用了MAP 算法，特定说话人的模型只需要较少的语音就可以训练出对应的高阶数模型。

4.3.2 基于 UBM-GMM 说话人确认的确认过程

在识别阶段，对于一段给定的语音（$X=\{X_1, X_2, \ldots, X_T\}$）,测试的语音特征参数分别计算与UBM 模型和目标说话人（Target）的似然度输出，两者相减，得到测试语音特征矢量X_t的似然度比，公式如下[49]：

$$\Lambda(X)=\frac{1}{T}\sum_{t=1}^{T}(\log p(X_t\mid\lambda_{tar})-\log p(X_t\mid\lambda_{UBM})) \tag{4.31}$$

X_t为一帧为一帧语音特征参数，T 为测试语音的总帧数，λ_{tar}为特定说话人的模型，λ_{UBM}为背景模型。

相应的似然比分数（帧似然比）为：

$$C(X_t)=\frac{p(X_t\mid\lambda_{tar})}{p(X_t\mid\lambda_{ubm})} \tag{4.32}$$

由于在UBM 中和Target 模型中都有刻画背景噪声的部分，而且由于Target 模型是由UBM 做说话人自适应得到的，因此在做自适应的过程中，只有UBM 中描述非特定人的语音分布的部分参数会得到修正，而UBM 中描述背景噪声的参数不会被修正，所以，对于测试语音中的噪声段，UBM 和Target 模型的输出十分接近，在相减后，可以很好的抑制背景噪声的影响。因此，UBM-GMM 系统作为目前与文本无关的说话人确认系统最流行的方法，与普通GMM 相比具有更好的性能以及噪声鲁棒性。图4.9为基于UBM-GMM 的说话人确认系统的识别框图。

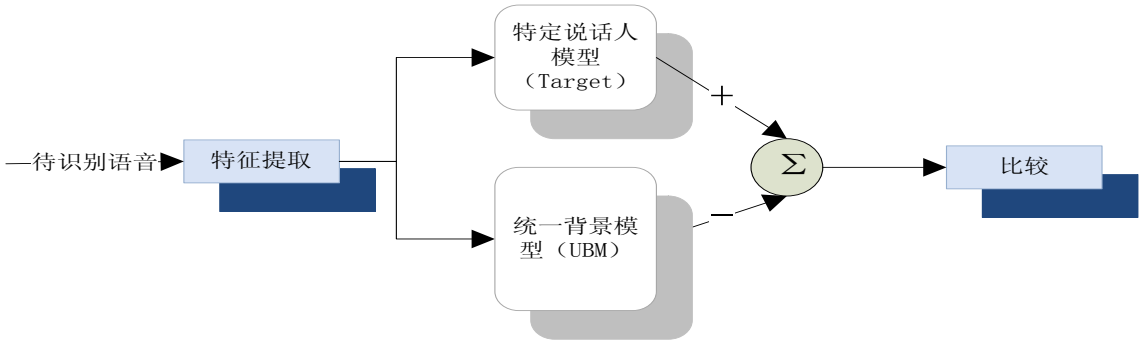


图4.9 识别过程的流程图

4.4 GMM 与 GMM—UBM 的性能比较

实验是所用的语音是TIMIT库里面的语音叠加上NOSIE-92里面的背景噪声。实验中，除了上面介绍到的预处理之外，需要对语音信号进行端点检测。测试者经常会出现说话不连续的情况，这样测试语音含有大量的静音(Silence)段严重影响确认性能。所以我们首先对语音信号进行端点检测(Endpoint Detection)，去除信号中的静音段和噪声(Noise)段。端点检测用来决定语音信号的开始和结束位置，错误的端点可能使语音被误认为静音或噪声，或者把它们误认为有效的语音。因此，端点检测的效果好坏直接影响着确认的准确率[48]。在比较时候，对GMM训练模型和对UBM的自适应的目标说话人带噪语音是相同的。GMM-UBM的混合度设定的是2048，对GMM模型的混合度设定的是128。比较两个系统的DET曲线，实验结果如图4.6所示，可见GMM—UBM的性能要远远好于GMM。

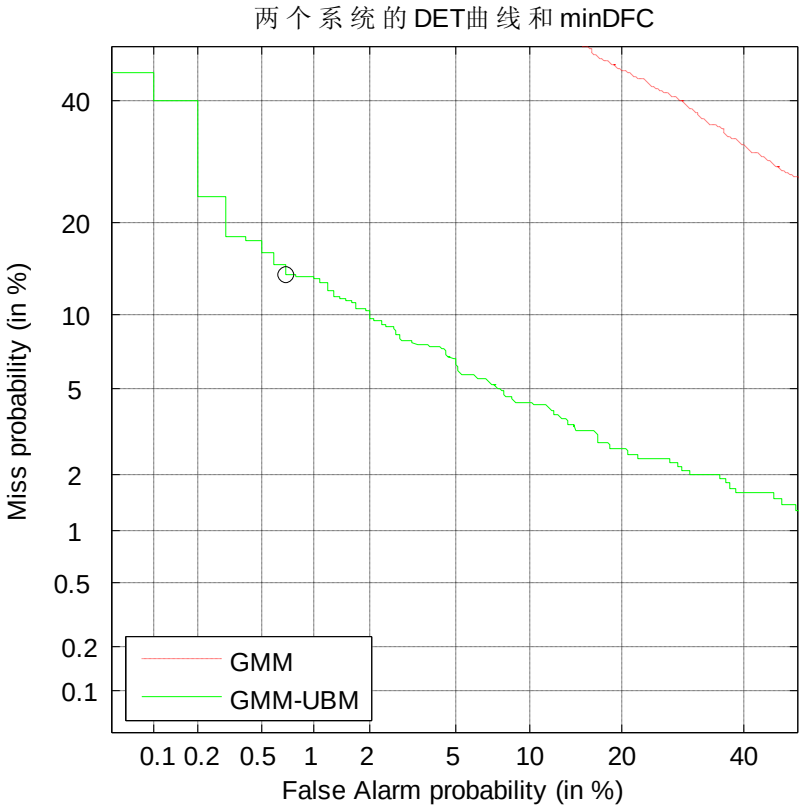


图4.6 GMM和GMM-UBM的性能比较

4.5 本章小结

本章介绍了GMM模型的和基于它的与文本无关的说话人确认系统，并提出了一个新的特征参数的提取方法来与GMM模型相融合，提高了系统的识别率。从模型方面引

入了UBM，以提高识别系统的鲁棒性。介绍了基于GMM-UBM 的与文本无关的说话人确认系统，并比较了两个系统的DET曲线，可以看出GMM-UBM相对于GMM模型的优势。

5 GMM 和 SVM 的混合模型

这章内容主要介绍GMM和SVM的混合模型，首先我们来了解一下SVM模型。

5.1 SVM 模型

支持向量机(Support Vector Machine, SVM)[38]是由Vapnik首先提出来的，Vapnik在他出版的《Statistical Learning Theory》[25]完整阐述了统计机器学习思想。这本书里面对统计机器和学习传统机器学习之间的存在区别的根本原因和本质进行讨论和说明，这个根本原因也就是在于统计机器学习能够精确的给出学习效果，能够解答需要的样本数等等一系列问题。与统计机器学习的精密思维相比，传统的机器学习的理论系统并不完善，还需要大量的经验和研究结果来支持，如果用传统的机器学习方法构造分类系统完全成了一种技巧可能在不同分类情况下效果会相差很大，缺乏指导和原则。尽管如此SVM还是得到了广泛的应用，他在现在中应用的意义也不容小觑。

5.1.1 研究意义

虽然 GMM 等统计模型在说话人识别领域已经取得了很好的成果，SVM 作为一种分类判决模型，其研究价值也不可小觑。相比于统计模型，SVM 是一种新的并且也十分可行的方法，在说话人确认模型中，他的判定方法与统计模型的判定方法截然不同。SVM 可以对统计模型做到很好的补充统，统计模型适合处理连续信号，SVM 适合分类；统计模型受最大似然准则的限制，找出特征参数的相同之处，把具有相同特征的参数归为一类，类别区分能力较弱，而 SVM 则是找出不同特征参数之间的不同之处，进行分类，能体现异类样本间的差异。SVM 在很多模式识别领域都取得了很好的成果，列如人脸识别、手写字体识别，这些识别对象的特点就是提取出来的特征向量是独立的。而语音信号确是连续的，根据对语音信号的处理我们可以探索 SVM 在连续输入量问题的解决能力。

SVM 是基于统计学理论的 VC 维（Vapnik Chervonenkis Dimension）理论和 SRM（Structural Risk Minimization）准则的。统计学习理论(Statistical Learning Theory,SLT)是一种专门研究有限样本情况下机器学习规律的理论。SRM、也就是结构风险最小化会在下面的内容中专门提出。至于 VC 维、这里简单介绍一下，VC 维的提出是为了在一定程度上能够起到度量函数类的作用，简单的来说就是他表示了问题的复杂程度，更直观的解释就是一个问题会随着 VC 维的变高而被理解为变得更加复杂。后面对 SVM 的

研究可以表明，SVM 解决问题的时候，和样本的维数是无关的，经过核函数的引入之后，他可以对高维数的参数进行分类。原因也就是因为 VC 维是 SVM 关注的主要对象。

5.1.2 基于风险最小的机器学习方法

机器学习的学习本质就是希望完成对问题真实模型的逼近，使期望风险能够最小化，一个模型的期望输出和这个模型实际的输出会存在差距的，但是真实模型却是无法得知的，我们需要选择一个与真是模型最大程度相似的假设（也就是我们得到了一个分类器），假设与问题真实解之间的误差（也就是实际输出与期望输出之间的误差），就叫做风险 R(w)，（更严格的说，误差的累积叫做风险），为了对这个问题经行简单的描述，在这里设定变量 x 和 y，他们之间存在着某种未知相互依赖的关系，也就是能把 x_i 映射到 y_i 上，机器学习就是想要对这一依赖关系进行估计[39]，也就是让他们能有规律的依从某一个未知的联合概率 F(x,y)，我们在这里把 x 视为是输入的语音的特征矢量，y 是期望的输出。机器学习的学习过程实现有一个一个最简单的方法，就是让学习机记住每一个训练指令，为了增加此新的特征矢量 X 进行进一步的分类，只需要在训练数据中找出与之相近的矢量，再把这部分相近的矢量进行相应分类。而这种方法仅适合在训练数据相对较少的情况，若数据样本很大、数据量繁杂多变的时候是不能满足要求的，此时需要在大的数据样本中采用一些映射方式，比如 X:→f(x,w)，将其转换到{f(x,w)}函数（也就是预测集）里面去寻找最优函数，一般为 $f(x,w_0)$，通过 $f(x,w_0)$来对应输入输出信号，使输入输出之间产生对应关系，进一步评估以达到风险值 R(w)最小。风险 R(w)可以表示为：

$$R(w)=\int L(y,f(x,w))dF(x,y) \tag{5.1}$$

其中{(f(x,w))}表示了预测集，w 代表了函数的广义参数。理论上来说尽可能多的训练数据学习得到分类器效果越好。训练的目标是尽量减少学习所得到的分类器的实际输出与期望的输出之间的误差

$$\min(y_i-f(x_i,w))R(w)=\int L(y,f(x,w))dF(x,y) \tag{5.2}$$

在我们选择了一个假设之后，我们所选择的假设与问题真实解之间差距因为真实模型的未知性同样也是无法得知的，我们需要用一些可以被确定并且调控的量来逼近这个差距。这时候就想到了已经知道的训练样本，训练样本是正确的数据。使用它作为真实的结果和在样本数据上的分类的结果与真实结果之间的差值来表示这个差距。这个差值叫做经验风险(Empirical Risk)，用 $R_{emp}(w)$来表示：

$$R_{emp}(w)=\frac{1}{N}\sum_{i}^{N}(y_i-f(x_i,w)) \tag{5.3}$$

从式子中可以看出经验风险的提出就是希望把数学期望简单的用算术平均来进行代替，由于机器学习的目的是希望风险最小的，所以以前的机器学习方法提出了经验风险最小化(Empirical Risk Minimization, ERM)之后并把它作为努力的目标，不过经过研究后发现虽然有很多的分类函数能够在原本的样本集上轻易达到百分之百的正确率，却在真实使用时候对未知数据分类时却的推广能力泛化能力却很差。此时的情况便是选择了一个 VC 维数足够高的也就是足够复杂的分类函数，这个分类函数的学习能力很强能够精确的记住每一个样本，但在面对需要对样本之外的数据进行分类的情况时却没办法得到正确的输出。其中一个典型的例子就是神经网络的过学习问题，对于小样本的情况，神经网络的学习能力太好以致于它能在学习过程中记住每个样本，这时 $R_{emp}(w)$虽然能很快收敛但对未知样本的预测能力很差，误差过于小而使预测错误率增大，推广能力下降。（这里所说的小样本并不是传统意义上认为的数量很少的样本，不论是对于什么算法，样本的增加带来差不多一直能带来更好的结果，这里之所以称为小样本是由于 SVM 算法要求的样本数与问题的复杂度比起来是相对比较少的。）这个时候我们再仔细的分析一下 ERM 准则就会发现，他是以有限的已知样本为前提的假设风险，使他适用的必要条件是经验风险要确实能够很大程度上的接近真实风险才可以，但实际研究结果表明它并不能达到这个条件，因为样本数相要大大小于现实世界要分类的样本数，ERM 准则只在这占很小比例的样本上做到没有误差，不能保证在更大比例的真实文本上也没有误差。训练样本的有限和学习机器设计不合理导致了过学习的现象。试图采用一个相对复杂的模型去拟合有限的样本使学习机的推广能力下降。

从上面所述的过学习问题上我们可以看出在有限样本下，仅仅用 ERM 来近似期望风险在实际的实现过程中是行不通的，所以统计学习引入了泛化误差界的概念来对经验风险进行修正。泛化误差界的概念是认为一个模型的真实风险应该由上面所提到的经验风险和一个定义为置信风险的两个部分所代表的内容来对其进行描述的，经验风险根据其定义可以认为它代表了分类器在已知的给定样本上的误差，但是由于它并不能完全描述真实风险，所以我们需要一个置信风险来对它进行修正，置信风险也就代表了分类器对在未知文本上分类的所得到的结果的可信度。但是遗憾的是，第二部分与真实风险一样是没有办法精确计算的，只能得到一个范围来计算增提误差的最大值。由于这个值是不能准确的计算出误差的值所以它被称为泛化误差界而不是泛化误差。

置信风险与两个量有关，一是样本数量，显然给定的样本数量越大，我们的学习结果越有可能正确，此时置信风险越小；二是分类函数的 VC 维，显然 VC 维越大，推广能力越差，置信风险会变大。我们用 Φ(n/h)来表示置信风险，那么泛化误差界的公式就为：

$$R(w) = R_{emp}(w) + \Phi(n/h) \qquad (5.4)$$

依据这个不等式我们把经验风险最小的概念转换成了经验风险与置信风险的和最小，也就是所说的结构风险最小的概念。因为 SVM 是一种努力追求于最小化结构风险的算法，那么可以认为它是在追求这两个和的最小化。

SVM 的建立与训练样本要趋于无穷大时才能达到理论上的最优的统计学习方法不同。SVM 把复杂问题简单化的解决方法是需要找到待解决问题的最优分界面(Optimal Separating Hyperplane)，对判别函数进行调整使得它可以最好的利用边界样本点的分类信息，通过输入输出关系的比较学习，寻找那些对分类有较好区分能力的数据，用这些数据构造出的分类器可以最大化类与类之间的间隔，能够在很好的解决小样本情况下的机器学习问题。SVM 本身的概念并不复杂，但其很强的扩展性决定了它很好的分类性。支持向量机具有简洁的数学形式、直观的几何解释和良好的泛化能力，避免了局部最优解，有效的克服了概率统计模型中因为维数过多的问题而带来的困难，近年来被应用于人脸、说话人、手写字等生物识别、金融工程、数据挖掘等方面。基本上，理论上来说 SVM 可以适用于任何涉及到模式识别的问题，并且能够取得很好的使用效果。SVM 使用于说话人识别时，所建立的 SVM 模型主要关注是否能够准确地描述目标说话人和其它说话人之间的差异和区别，以便于就算遇到最小的分类误差模型依然会有很稳定、有很好的泛化能力。说话人确认是一个二元分类的问题，只需要得到是或否的输出，所以作为一种有区分性的辨别模型，SVM 很适合解决说话人确认这样的问题。

下面我们来介绍 SVM 的判决，SVM 方法是从两类模式识别时线性可分情况下的最优分类面提(Optimal Separating Hyperplane)出的。它们之间的距离叫做分类间隔(margin)。

5.1.3 分类间隔最大

SVM不仅仅是应用于平面内点的分类问题导致了它复杂性。SVM的一般将所有待分类的点映射到一个高维空间，然后在高维空间中找到一个能将这些点分开的超平面，这在理论上是被完全证明了是成立的，而且在实际计算中的也有可行性。不过在SVM中，仅仅找到超平面是不够的，因为在通常的情况下，满足条件的超平面的个数不是唯一的。SVM需要的是利用这些超平面，找到这两类点之间的最大间隔。为什么要找到最大间隔呢？我想这与SVM的推广能力有关，因为分类间隔越大，对于未知点的判断会越准确，也可以说是最大分类间隔决定了期望风险，总结起来就是：SVM要求分类间隔最大，实际上是对推广能力的控制。

为了使得分类间隔最大，需要找到两条相互平行的直线，每条直线都能够无差错的将两类数据分开。然后保持它们之间的平行关系不变，对其进行旋转、平移使得两条直线的间距尽可能的远，同时保证每条直线对两类数据的无差错分类能力。这样得到的两

条直线(分类超平面)就定义了两类之间的间隔，称为margin。通过适当的旋转、平移等变换，使得margin 最大，这样得到的分类线位于margin 的正中间，而此时这两条直线就被称为最优分类超平面。使分类间隔最大，实际上是对推广能力的控制。分类间隔越大，则推广能力越好，这是SVM 的核心思想之一[41]。

存在着一个超平面在线性可分的情况下可以使使得训练样本完全可分，表述如下：

$$w\cdot x+b=0 \tag{5.5}$$

它的推导过程如下：

对于线性可分的样本集，若分类线能将两类训练样本分开，则有：

$$w\cdot x_i+b>0 \qquad y_i=1 \tag{5.6}$$

$$w\cdot x_i+b<0 \qquad y_i=-1 \tag{5.7}$$

通过适当的调整w 和b ，可以将上式改写为：

$$w\cdot x_i+b\geq 1 \qquad y_i=1 \tag{5.8}$$

$$w\cdot x_i+b\leq 0 \qquad y_i=-1 \tag{5.9}$$

或将其归一化为：

$$y_i[(w\cdot x_i)+1]-1\geq 0 \qquad i=1,2,\cdots,n \tag{5.10}$$

显然，位于最优超平面上的样本的分类器输出值将为+1 或-1。由于支持向量机的学习目标是分类间隔 margin 最大，因此首先要解决如何求分类间隔的问题。事实上可以通过分别位于分类超平面上的一对样本求得分类间隔 margin，即

$$d(w,b)=\frac{w\cdot(x_2-x_1)}{\|w\|}=\frac{2}{\|w\|} \tag{5.11}$$

给“最优超平面”以下定义做说明：在空间中存在一个平面，这个平面可让一类数据和超平面距离最近的向量但距离最大的平面，称此超平面为最优超平面

$$\min\Phi(w)=\frac{\|w\|^2}{2} \tag{5.12}$$

满足最小化的条件且满足 $y_i(w\cdot x_i+b)\geq 1 \qquad i=1,2,\cdots,n$ 的分类面就叫做最优分类面

n 是样本的总个数。w 可通过转化为对偶问题及利用拉格朗日（Lagrange）乘子等数学方法求得：

$$\sum_{i=1}^{n}y_i\alpha_i=0 \text{和} \alpha_i\geq 0 \quad \text{i=1,2，…，n} \tag{5.13}$$

求解下列函数的最大值:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i,x_j) \tag{5.14}$$

a_i 为原问题中与每个约束条件对应的拉格朗日乘子。这是一个不等式约束下二次函数寻

优的问题，存在唯一解。容易证明，解中将只有一部分 a_i 不为零，对应的样本就是支持向量。解上述问题后得到的最优分类函数是

$$f(x)=\mathrm{sgn}\{(w\cdot x)+b\}=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i^{*}y_i(x_i\cdot x)+b^{*}\} \qquad (5.15)$$

式中的求和实际上只对支持向量进行。b^{*}是分类阈值，可以用一个支持向量求得，或通过两类中任意一对支持向量取中值求得。

如果存在一个分类函数，它可以将两类样本完全分开。一般的，如果一个线性函数能够将样本完全正确的分开，就称这些数据是线性可分的，否则称为非线性可分的。

5.1.4 线性判决边界

假定一个对于线性可分的样本集及其映射

$$(x_i \rightarrow y_i) \quad i=1,...,n \quad x\in Rd \quad y\in\{+1,-1\} \qquad (5.16)$$

其中 x 代表了训练样本，y 代表了这个训练样本所属的类。在 SVM 分类过程中,并不重视分类器的输出值的大小，重要的他的值是为“正”还是为“负”。也就是说当 f(x,w) 大于零时，将其归于+1 类，反之时则将其归于-1 类。也就是“硬输出”，f(x,w)定义了将两类区分开的判决边界[40]。

我们在这里讨论最简单的线性关系，即

$$f(x,w)=x\cdot w+b \qquad (5.17)$$

式中，x 表示输入向量，w 表示权系数向量，b 表示分类阈值。

5.1.5 非线性判决边界

以上对线性可分时间的解决方案问题做了分析，然而这仅仅局限于理论上，实际运用过程中需要解决的问题中各种繁杂的情况是线性且不可分的，这就导致使之前提到的优化问题将变得无法解答。对于线性不可分的情况，采用将样本X 映射到高维空间Y，并在高维空间中采用线性方法进行处理。只要核函数满足Mercer 条件[44]，它就对应某一空间中的内积，因此只要在最优分类面上采用适当的内积函数就可以实现这种线性不可分的分类问题。

高维空间映射：由于非线性判决边界的求取以及描述都存在相当的问题，很难解决，并且很难对其进行参数化调整，基于实际情况的不可变因素，使用一种合适的解决方法就是找寻一个线性的判决边界，通过此边界来描述此种非线性情况下两个或者几个类样本之间的分类面。因为在原始特征的空间中(d维)，这几类样本之间的判决边界是一种非线性的复杂解决问题，所以不能直接由一个或者多个线性的判决边界来届时两类样本

之间的非线性边界之间的问题，然而这些问题是可以通过某种非线性的特征变换来解决的，具体方法是将空间维度提高，将原始的非线性问题转化至高维空间中的线性问题去解决，进而在这个高维空间中创建广义的最优超平面，通过这种方式可以就可以原始特征空间中非线性问题（转化为高维空间的线性问题）。

为了使非线性可分的数据映射到一个高维空间中从而使其线性可分，需要引入下面两个概念：松弛因子和核函数。

（1） 核函数[21]，几个实例的解决方案中可以看到，支持向量机的在此过程中的一个很重要的特点和优点，也是解决问题的关键点，那就是能够将在原始空间里存在的不可分问题通过特定的非线性变换，然后将此转换至一个高维空间中，在高维空间中，此问题就不在是非线性的，而是一个可以分割的线性问题，然后在变换的高维空间中寻找最优分类面，这样就能够很容易的用一个简单的线性分类器来解决其最复杂的分类问题。但是在这个过程中，怎么变换以及变换至哪个空间以及哪个空间是我们可以解决问题的空间等等问题都是由核函数所决定。综上所述，在支持向量机的理论体系之中，核函数的地位是非常非常重要的。然而由于数据分布的形式事先是并不所知，怎么将这种未知分布的而且不可分割的数据变转换至一个未知空间中的可分数据，就成为支持向量机需要解决的第一个问题。当然这种变换可能比较复杂，那么这种思路在一般情况下很难实现。但在上面的对偶问题里面，无论是寻优目标函数还是寻求分类函数都只需涉及训练样本之间的内积运算方式(x_i x_j)，没有单独的 x_i 出现在其中。所以若能够找到一个函数 K，并满足 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$，那么这样在高维空间中，问题的实质实只需进行内积运算问题。而且内积运算又是可以利用原空间的函数帮其到达，所以我们可以不需要知道变换 φ 的具体表达式。由泛函的有关理论“只要一种核函数满足 Mercer 条件，那么它就对应某一变换空间中的内积”。因此这个事实说明对构造支持向量机及其重要程度，由于当特征空间的维数巨大的情况下，我们无法想象直接计算其内积的计算复杂度。

所以在最优分类面中采用适当的核函数就可以实现某一种非线性变换后的线性分类，那么此种分类问题迎刃而解，而且计算的复杂度并未增加。只是目标函数变为：

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{5.18}$$

相应的分类函数为：

$$f(x)=\mathrm{sgn}\{(w\cdot x)+b\}=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i^{*} y_i K(x_i,x_j)+b^{*}\} \tag{5.19}$$

我们在构造判别的函数的时候，并不是对输入空间的样本直接进行非线性变换后在特征空间中去求解，而是需要先在输入空间中做某种比较后，再对比较后的结果做非线性变换。那么这大量的工作就是在输入空间去完成计算操作。简单的讲，支持向量机就

是首先把非线性变换后将输入数据变换到一个高维的特征空间中去运行，而且在这个高维空间的特征空间中去寻求广义的最优分类面。SVM 的分类函数在形式上与一个单隐层的神经网络类似，而且输出是隐层节点的相关的线性组合，其中每一个隐层节点对应的一个支持向量。如同前面所介绍的那样，引入核函数的概念的目的是将原始空间中的不可分数据转换为高维空间中的可分数据。现阶段看来，核函数的选择通常都是在训练之前根据经验来选定合适的核函数再去进行下一步的工作，包括核函数的类型以和相应的核函数的相关参数。对于不同的说话者，虽然每个说话人的训练数据的分布形式有很大的差异，但不同的说话人所采用的核函数模型却是一样的。当然理想情况下的这种转换应该与原始空间中的数据分布有密切关系。所以近年来开展了一些关于与数据有关的核函数如何选择的研究课题。这些研究的实质是把核函数作为一个模型来进行研究处理，换句话说，核函数本身就是一个有区分性的模型。所以每个说话人的数据分布不同就会导致得到不同的核函数类型。因为这个核函数是在训练后得出的，那么就不存在核函数类型的选择问题，这样便可较为真实的反映了此种不可分的训练数据在原始空间的分布关系，同时也可以避免了核函数中参数的选择问题。但是，这样也会带来另外的问题。因为这个核函数是训练得到，与一切其他有区分性的模型训练问题一样。首先，这样训练非常依赖于训练数据的选择问题。当训练样本的分布和测试环境的分布存在较大差异时，这种训练所得出的核函数由于注重关注于对训练环境的描述，会导致推广性、即泛化能力的下降。其二，因为核函数需要事先训练，这样也会导致额外的运算开销。综上目前的一些主流做法依然是事先确定好核函数。

表 5.1 几种常用的核函数

核函数类型	核函数
线性核函数（Linear）	$K(x_i,x_j)=(x_i,x_j)$
多项式核函数（Polynomial）	$K(x_i,x_j)=[\gamma(x_i,x_j)+\eta]^d$
辐射基核函数(RBF)	$K(x_i,x_j)=\exp\{-\frac{\gamma\mid x_i-x_j\mid^2}{\delta^2}\}$
Sigmoid 函数	$K(x_i,x_j)=\tanh[\gamma(x_i,x_j)+\eta]$

经过核函数的讨论之后这里我们可以联想非线性判决边界到另一个问题，那就是如果使用核函数向高维空间映射后，问题仍然是线性不可分的，要如何解决，这个问题的提出就引出了另一个参数：松弛变量。

（2） 松弛因子，现在假设我们已经把一个本来线性不可分的问题映射到高维空间而变成了线性可分得向量，如图5.1所示。

假设有另一个训练集，比这个多了一个样本点，映射之后，如图 5.2 所示。

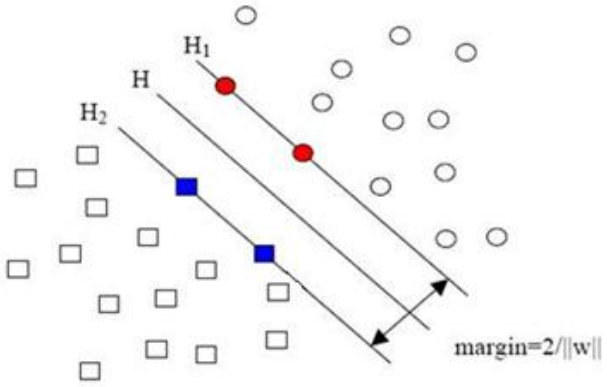


图5.1 一般的线性可分向量

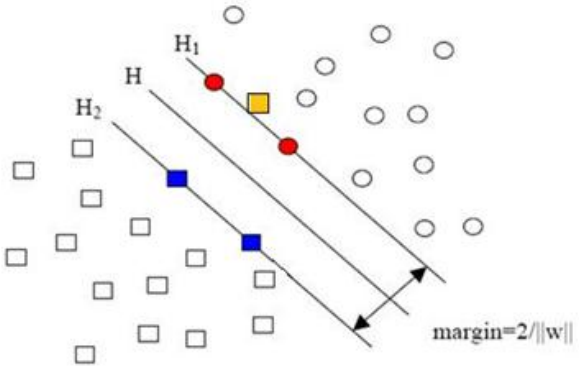


图5.2 多了一个样本点的情况

这单独的一个样本，是的原本线性使得原本线性可分的问题变成了线性不可分的。这样类似的问题（仅有少数点线性不可分）叫做“近似线性可分”的问题。本来有大量的数据都是符合原来的分布的，我们是否应该为了这一点而对原来的规则进行修改，这一点能否代表分类没考虑到的样本。其实它可能只是一个噪声，那么如果忽略了这点，仍然使用原来的分类器，其效果丝毫不受影响。我们不希望因为这一点噪声造成整个样本的最优分界面无解，所以我们需要系统对这种对噪声存在容错性，实现方法也就是对原来的硬性的阈值加一个松弛变量 ξ。下面来对它的实现方法进行讨论。

在实际需要解决的问题中，很多情况是线性不可分的，因此，优化问题将变得无解。因此，若训练样本集是线性不可分的，或事先不知道其是否可分，引入非负变量 ξ_i,i=1, ··; n，使得：

$$w \cdot x_i + b \geq 1 - \xi_i \qquad y_i = 1 \tag{5.20}$$

$$w \cdot x_i + b \leq \xi_i - 1 \qquad y_i = -1 \tag{5.21}$$

引入松弛因子ξ_i将允许某些样本落在分类间隔区域内，甚至允许出现分类错误。这样实际上是在一个允许出错的情况下来解决线性不可分情况下的样本分类问题。在这里ξ_i是非负的松弛变量，可以看作是对训练样本关于分类超平面的偏差的惩罚。对于一个落在决策域的右边区域但是仍然位于分离区域内的样本点，ξ_i的值将小于1。而对于一个落在错误决策域内的样本点将给以更大的惩罚，使ξ_i >1。而ξ_i=0时就是线性可分情况，因此总的分类错误小于所以ξ_i的和。在目标函数中加入一项对错误分类进行惩罚的项，折衷考虑到最大分类间隔和最小分类错误之间的平衡，这样就得到的线性不可分情况下的支持向量机，则相应的分类面优化问题为：

$$\min\{\frac{1}{2}\|w\|^2\} + C\sum_i L(\xi_i) \tag{5.22}$$

且

$$y_i[(w_i \cdot x_i + b)] \geq 1 - \xi_i \qquad \xi_i > 0 \qquad i = 1,2,\cdots,n \tag{5.23}$$

其中C>0 是自定义的惩罚系数，用来控制样本偏差与机器泛化能力之间的平衡。在最小化上述目标函数的过程中，对于C和ξ产生了一个相反的影响。对于较小的C值，ξ影响较大，也就是说，得到的解在最小化错误率时倾向于得到一个较小的裕度。而对较大的C值，ξ在最小化中的影响就较小，也就是说，此时可以容忍较大的分类错误从而使得分离裕度较大。在实际应用中，对于C值的选取往往必须通过实验的方法来定，因为它可能具有不止一个的“最优”值。L 则是对分类错误进行惩罚的损失函数。它的拉格朗日对偶问题为：

$$\max\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j)\} \tag{5.24}$$

且

$$\sum_{i=1}^{n} y_i\alpha_i = 0 \qquad 0 \leq \alpha_i \leq C \qquad i = 1,2,\cdots,n \tag{5.25}$$

和

$$w = \sum_{i=1}^{n} y_i\alpha_i x_i \tag{5.26}$$

下面我们用libsvm工具箱对工具箱自带的数据heart_scale进行SVM的分类并画出它的ROC曲线，如图4.4和图4.5所示。

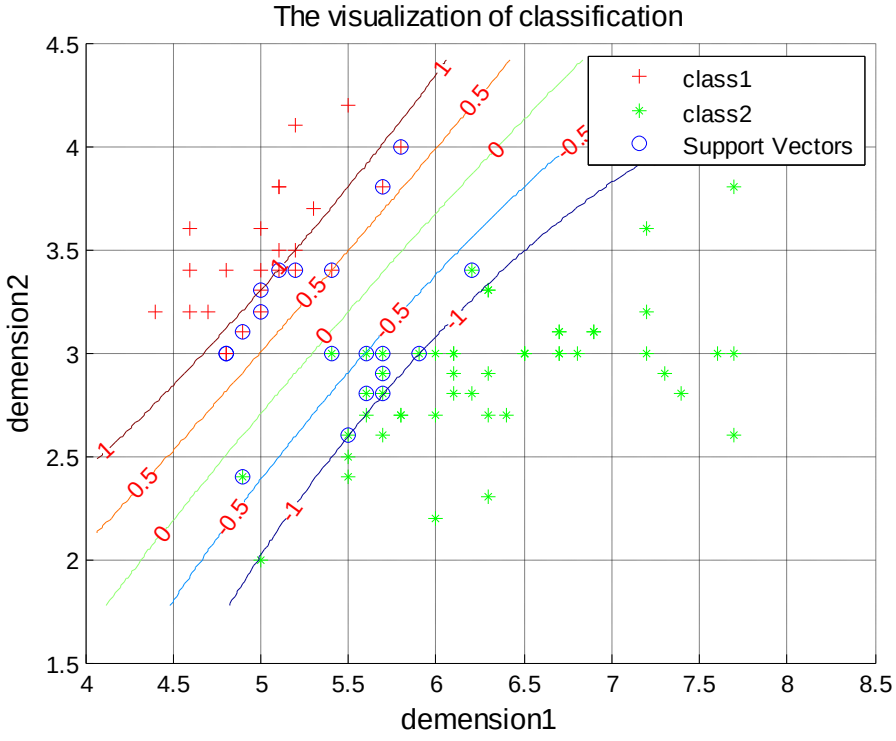


图5.4 SVM对heart_scale数据进行分类

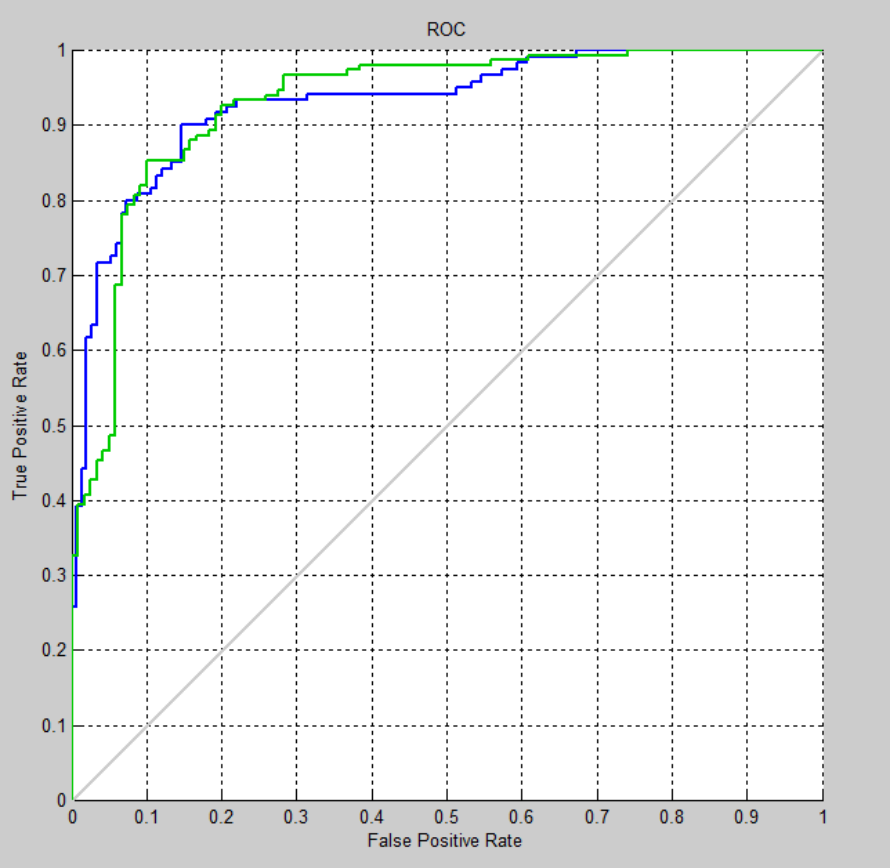
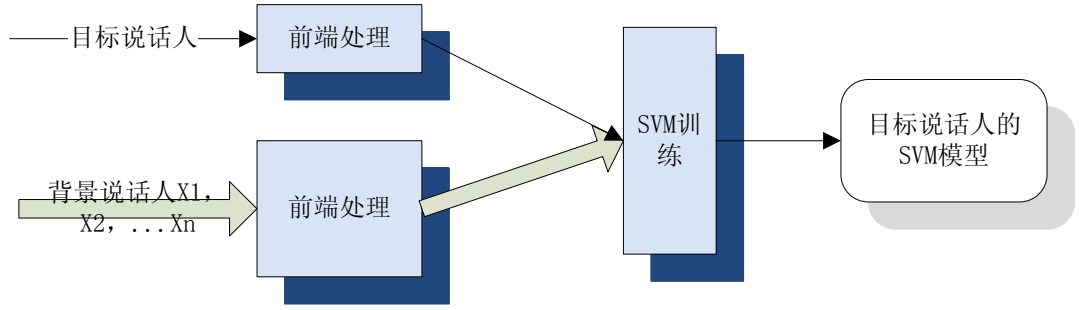


图5.5 SVM对heart_scale数据进行分类得到的ROC曲线

5.2 SVM 在说话人确认中的应用

基于支持向量机的说话人确认系统的框图如图5.6所示。系统分为训练(注册)和识别（登录）两个阶段。在训练阶段，系统由说话人的注册语音训练得到该说话人的支持向量机说话人模型，标记后存贮在模型库中；在识别阶段，测试语音在给出的同时会附带一个身份说明。此语音信号在经过前端处理转换为特征参数序列之后，将按照其所声明的身份说明从模型库中调出所声明的说话人模型，并与之进行匹配，给出相应的得分。然后这个得分将与确认阈值进行比较，最终获得确认结果(接受或拒绝)。



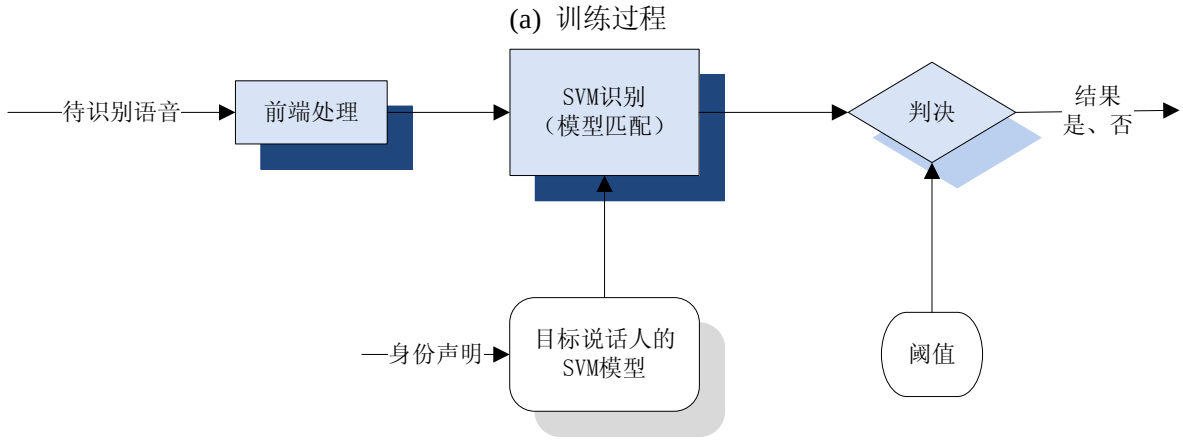


图5.6 基于SVM模型的说话人确认系统的训练框图

支持向量机作为一种辨别模型，其模型的训练是一种有区分性的训练，即其模型训练的目标函数是有区分性的，同时其模型训练也需要正反两类数据，分别是目标说话人(Target Speaker)和冒认说话人(Imposter Speaker)的训练语音数据。冒认者一般选择与所有目标说话人没有交集的语音库，我们称其为背景说话人(Background Speaker)。在对某一个目标说话人进行SVM 建模时，目标说话人特征一般标记为+1 类，而上百的背景说话人特征标记为-1 类。SVM 实际上是刻画真实说话人和冒认者之间的区别，其分类面由真实说话人和冒认者的数据决定。SVM 模型训练的目标是通过相应的算法，构造出一个合适的分类面，能够很好地将目标说话人和冒认说话人有效地区分开来，同时还要对未知的测试环境具有较好的泛化能力。在识别阶段，如果是未知说话人特征与对应的说话人SVM 模型分类面十分接近，那么获得的分数就大。如果超过阈值则未知语音判别为相应的说话人的，反之，若分数低于阈值就判别为冒认者。

5.3 GMM-SVM 系统

在有些文献中[22、44]，直接使用语音特征作为 SVM 训练和识别的参数。但是在与文本无关的说话人确认中的模型建立中，需要较长的训练语音来体现出说话人的个人性信息和大量背景说话人的语音来解决测试语音与模型不匹配的情况。所以将语音信号的参数直接用于 SVM 时，会面临特征矢量样本集数量大、两类数据混叠严重而导致的 SVM 说话人模型训练的困难。在与文本无关的说话人确认中，常用的声道参数无法完全分开说话人个性特征信息和语义内容的信息。将这种声道参数作为 SVM 说话人模型的训练参数往往会影响系统的识别性能。参数会在很大程度上影响 SVM 的识别性能，所以在 SVM 模型中输入的参数需要是经过塞选合适的样本以减小对分类效果的影响，具有较少的支持向量、较小的模型存储量、较好的泛化能力以及较高的模型测试效率的支持向

量机模型，是支持向量机模型训练参数选择的一个理想的目标。显然，在语音识别和说话人识别中常用的声道参数不能满足支持向量机模型训练实时性以及分类效果的要求，有必要对声道参数做进一步的处理，从而得到所需的特征参数。所以提出 GMM 模型和 SVM 模型的混合模型。

GMM模型和SVM模型，因为一个是属于概率统计的模型，这类模型基于贝叶斯判决理论，它将模式识别中的归类问题转换为对数据分布的估计问题，采用概率密度估计的方法来对说话人特征的统计分布进行尽可能精确的描述，从而将复杂的语音训练、匹配的问题分解为模型表达形式的选择、参数的训练、以及概率的计算等等子问题。另一个是基于判别分类模型的方法。如果说概率统计只专注于对同一类数据的统计分布进行细致的描述，判别分类模型则是对不同类之间的分类面进行刻画，使得不同类之间的分类误差尽可能的小。两类方法虽然各有特点，识别性能却相当。概率统计模型的特点是可以充分统计数据的分布情况，找出同类数据的相似的特性；而判别分类的模型不同类数据之间的最优分类面，找出不同类数据之间的差异，对内部的特征分布则是不关心的。文献[40]研究表明了两者在同一训练数据的情况下，错误识别却有很大的不同，可以看出两种模型的互补性。判决分类模型有概率统计模型所没有的优点，反之亦然。如果通过算法下使两类算法有机结合，结合两者的优点，系统的识别性能可以有所提高。

我们来对这种方法来进行说话人确认进行实验。训练目标说话人和背景说话人（这里选了9个背景说话人）语音的GMM，每个说话人有64个样本，用其中32个样本作为训练语音，另外32个作为测试语音。依旧使用libsvm工具箱，将目标说话人的样本标记为1，背景说话人标记为-1，对模型进行训练，然后用训练得到的模型对剩下的样本进行测试，得到结果如下：

```
optimization finished, #iter = 359
nu = 0.149029
obj = -50.816907, rho = 0.849972
nSV = 187, nBSV = 11
Total nSV = 187
Accuracy = 96.5625% (309/320) (classification)
```

证明了两种方法结合的可行性。

概率生成模型（G）和概率判别模型（D）的融合一般分为三类情况：

（1） G 和 D 并列。例如文献[41]提出 GMM/SVM，经过试验确定两者在融合方法中所占权重的比列之后，取得折中方案。

（2） D 内嵌到 G 中。例如文献[42]]提出 SVM/HMM 的方法，用 SVM 的输出替代原来 HMM 的概率的输出。

（3） G 内嵌到 D 中。例如文献[43]提出的用核函数将 HMM 的得分映射到得分空间，再用 SVM 在此空间内进行训练和识别。

5.3.1 SVM 的概率输出

支持向量机作为分类器，并不能很好的处理语音信号这种连续信号，为此可以引入 GMM 等统计模型。称 GMM 为生成模型，SVM 为判决模型。已知 SVM 在的输出是二进制的，只输出是或者否，也就是“硬输出”，对于单个测试向量的分类来说，这样的输出相对容易处理，但是对于一个连续信号而言，这样的输出会掩盖许多信息。为了更好的融合两种模型。需要将 SVM 的结果转换为后验概率，表示为“软输出”，我们将 SVM 的输出通过 Sigmoid 转化为后验概率。从“硬输出”变为“软输出”。

已知对于输入向量 x 支持向量机的输出格式为：

$$y = sign(f(x)) \tag{5.27}$$

其中：

$$f(x) = w^T x + b \tag{5.28}$$

基于此，通过 sigmoid 函数给出了 SVM 的概率输出形式

$$P(C_{+1} \mid x) = \frac{1}{1+e^{-f(x)}} \tag{5.29}$$

和

$$P(C_{-1} \mid x) = \frac{1}{1+e^{f(x)}} \tag{5.30}$$

显然，处在分类面上的点对应+1 和-1 类的概率都是 0.5。

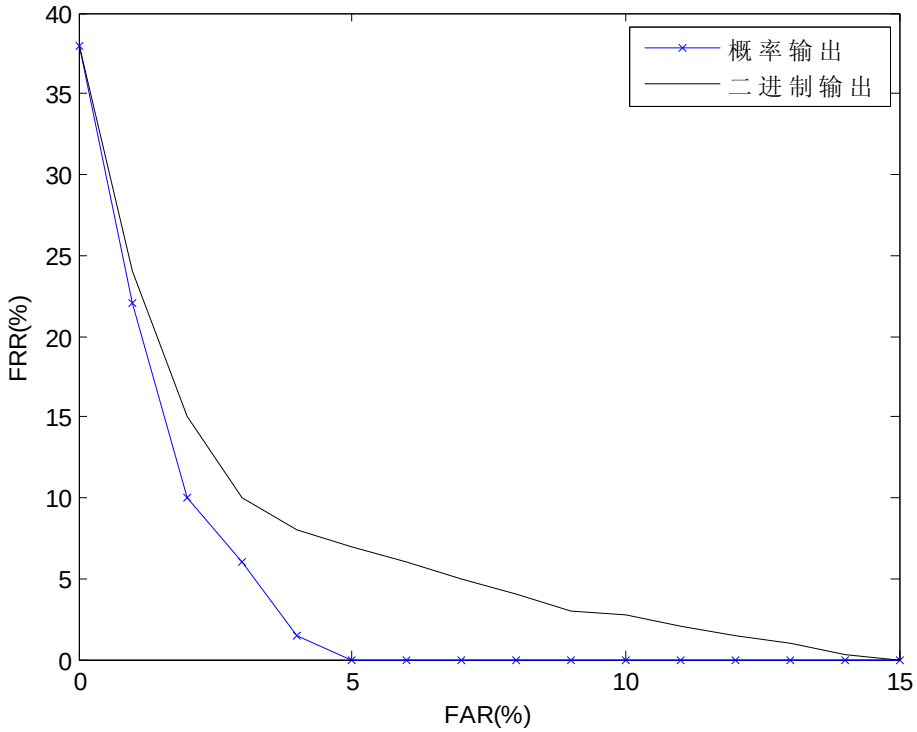


图5.7 概率输出和二进制输出的比较

充分结合 GMM 的鲁棒性和 SVM 的区分性，可以提高与文本无关的说话人确认系统的识别性能。GMM-SVM 系统的思想就是利用高斯混合模型对提取出来的特征参数参数作进一步的处理，得到数据量小却能够体现出说话人特征分布、尽可能去掉语义信息干扰、以及具有较好的区分性的特征参数。首先，高斯混合模型通过概率密度估计的方式，用若干个高斯概率密度函数的线性加权组合来对复杂的数据分布进行准确地描述，从而只需要相对较少的参数就能实现对大量数据的刻画，也即这些相对较少的参数可以起到代表大量数据的作用。GMM 模型只刻画语音特征参数的静态分布而不刻画语音的时序过程。由于不同说话人的语音特征的静态分布也是不相同的。因此，通过比较不同说话人的 GMM 模型，就可以区分不同的说话人。可以认为 GMM 描述了去掉了语音中的语义信息得到了说话人的个性特征参数。GMM-SVM 就是在其中前端处理使用了由某个说话人的声道特征参数进行训练得到一个对该说话人特征分布进行抽象描述的 GMM 模型的参数，再采用 GMM 的模型参数作为 SVM 说话人模型的训练数据进行判决。假设统一背景高斯混合模型 UBM 为：

$$g(x)=\sum_{i=1}^{N} w_i N(x, m_i, \Sigma_i) \tag{5.31}$$

对于这个统一背景高斯混合，进行最大后验概率(MAP)自适应得到与特定说话人的GMM，也就是上面所介绍到的GMM-UMB的自适应过程。这样经过自适应后所有特定

说话人的方差一样，反映了不同的说话人特征在特征空间的相对位置均值却不相同，均值位置正是不同说话人之间最有区别性的地方。本GMM-SVM系统就是采用GMM模型的均值。是参与支持向量机模型训练的样本个数而不再是说话人的声道特征矢量个数L，而是GMM模型的混合度个数N(N<<L)，这样便可以大大减少训练样本的数量；再者，GMM模型的参数是去掉了说话人特征矢量中的语义信息，从而突出了说话人的个性信息，使之更适合于与文本无关的说话人确认。另外，为了加大不同说话人之间的区分度，我们通过GMM模型对说话人特征矢量的抽象描述，使SVM模型训练能够更迅速的收敛，这样得到的支持向量个数较少，而且具有更好的分类效果。根据之前支持向量机的介绍我们可知，支持向量机在测试阶段的时候，实际上是将测试语音参数变换到核函数与支持向量所决定的高维空间中，然后在这个高维度的空间中计算该测试语音参数和分类面之间的距离大小，一般我们把这个距离一般称为是支持向量机的软输出[45]，如下表示：

$$f(x)=(w,x)+b=\sum_{i=1}^{n}a_i^{*}y_iK(x_i,x)+b^{*} \tag{5.32}$$

在我们的GMM-SVM系统中，这个距离指的是测试语音GMM模型的各个均值在变换空间中与由支持向量机所决定的分类面之间的距离测度。常用的输出评分定义是取个均值在整段SVM 模型上所获得分数的平均值，即：

$$S=\frac{1}{N}\sum_{i=1}^{N}f(x_i) \tag{5.33}$$

式中为N为用来表示测试语音特征分布的的GMM的个数。评分反映了测试语音的特征在变换空间中与分类面之间的平均距离。在理想情况下，如果测试样本来自于目标说话人，就输出正的软输出，反正，就输出负数。

5.4 基于 GMM super vector 线性核函数的 SVM 用于说话人确认系统

对于一个特定的说话人，利用最大后验概率（MAP）自适应对统一背景高斯混合（UBM）经行参数调整得到该特定说话人的高斯混合模型（GMM），把其中每个高斯的均值按序号依次连接起来所形成的超向量含有特定说话人的大量个性特征信息，SVM 分类就是在这个超向量上进行的。而把这些超向量看成不同类别，GMM 支持向量一般要形成一个上万维的向量，这可以看成将一句语音映射到一个高维空间上。这就是主动将 GMM 的 N 个均值映射到了高维空间上，这个概念与 SVM 的序列核的概念十分吻合[20]。序列核中一个核 K(x ,y)被用来两完成段语音 x 和 y 的比较。因为这里的核函数是一种满足 Mercer 条件的线性核函数，Mercer 条件保证了 margin 的有效性、保证了 SVM 的最优性，可以表示为：

$$K(x,y)=b(x)^{t}b(y) \tag{5.34}$$

式中，b(x)就要将 x 所在的输入量映射到一个高维空间。

GMM super vector 可以完成将语音映射到另外一段语音 b(y)上。

GMM 超向量（GMM super vector）是 Campbell 在文献[23、24]中提出的，他是一个很有效的线性核函数。

高斯模型问的距离(Distance Between GMMs)[20]，我们用MAP自适应方法对两段语音utt_a和utt_b进行训练得到各自对应的GMM模型。为了能区分两个GMM模型的概率分布，我们使用了KL散度(Kullback-Leibler divergence)来定义他们的距离。

$$D(g_a \| g_b)=\int_{R^n} g_a(x)\log\left(\frac{g_a(x)}{g_b(x)}\right)dx \tag{5.35}$$

不幸的是 KL 散度并不满足 Mercer 条件，所以它很难（不是不可能）被运用与 SVM 模型。为了对 KL 进行约束，我们使用对数和的不等式，得到一个近似于 KL 的方法：

$$D(g_a \| g_b)\le\sum_{i=1}^{N} w_i D(N(\cdot;m_i^{a};\Sigma_i)) \| D(N(\cdot;m_i^{b};\Sigma_i)) \tag{5.36}$$

式中利用 m^a 和 m^b 对原来的方法经行了改变，又因为它以 UBM-GMM 系统为前提，利用他训练的特点把上式近似表示为：

$$d(m^a,m^b)=\frac{1}{2}\sum_{i=1}^{N} w_i(m_i^{a}-m_i^{b})\sum\nolimits_i^{-1}(m_i^{a}-m_i^{b}) \tag{5.37}$$

结合上面的式子可得

$$0\le D(g_a \| g_b)\le d(m_a,m_b) \tag{5.38}$$

KL 散度会随着 m_a 和 m_b 之间的距离的变化而变化，所以可以用这个距离来衡量。

这个距离具有对称性。它已经成功被应用到说话人的聚类等问题。相应的内积函数，也就是我们寻找的核函数可以从该距离中推算出来：

$$\begin{aligned} K(utt_a,utt_b)&=\sum_{i=1}^{N} w_i m_i^{a}\Sigma^{-1}m_i^{b} \\ &=\sum_{i=1}^{N}(\sqrt{w_i}\Sigma^{-\frac{1}{2}}m_i^{a})^{t}(\sqrt{w_i}\Sigma^{-\frac{1}{2}}m_i^{b})^{t} \end{aligned} \tag{5.39}$$

这个一个线性的，且涉及到 GMM 超向量简单的对角线尺度的和向量，这个核在 GMM 超向量空间上进行了简单的线性加权，因为它是线性的，所以它可以满足 Mercer 条件。把 GMM super-vector 用于 SVM 系统的前端处理部分，作为 SVM 的参数输入，核函数就是 $K(utt_a,utt_b)$。我们实现三个系统并通过 DET 曲线来比较三个系统的识别性能。如图 5.8 所示。

5.5 本章总结

本章首先介绍了 SVM 系统，并对它运用于说话人确认系统进行了理论上的介绍和实验上的实现。根据它在说话人确认系统运用上的缺陷，介绍了 SVM 和 GMM 相结合的方法。从本章的不同的实验结果可以看出，SVM 和 GMM 两种模型的结合可以在一定程度上提高系统的性能，尤其是在引入了超向量之后，系统的性能得到了很好的提高。

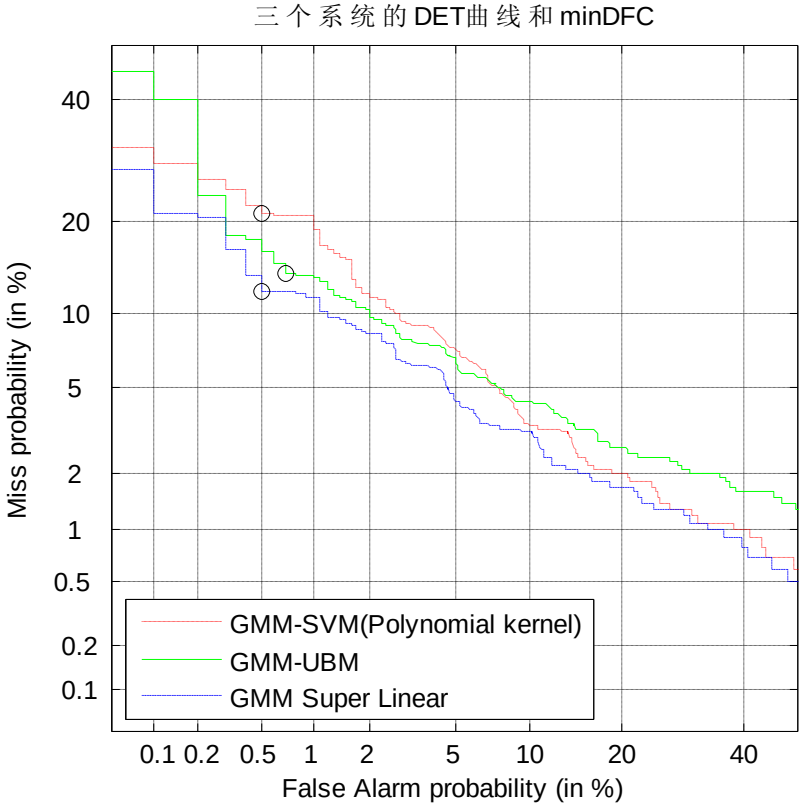


图5.8 三个系统的DET曲线的比较

6 总结与展望

6.1 工作总结

本文对复杂环境下对特定说话人的识别进行了研究，主要研究了在噪声环境下的与文本无关的说话人确认。为了提高识别系统的噪声鲁棒技术，从基于识别系统的特征和模型两个方面对噪声对系统参数的影响进行消除。在基于特征的噪声鲁棒技术中介绍了PCA、RASTA 滤波和特征弯曲等优秀的噪声鲁棒技术，并对他们进行了实现和分析，将这些当前最为有效的一些参数鲁棒性算法融入到系统的各种参数提取中，使得基本的系统都有良好的识别性能。并在介绍 GMM 模型时提出了一种新的抗噪的特征参数，实验的结果证明了这种特征参数的有效性。在基于模型的噪声鲁棒技术中，介绍了UBM-GMM 模型和 SVM 和 GMM 混合的模型。UBM-GMM 由于其较好的鲁棒性成为了与文本无关的说话人识别中的主流技术，可以从实验得出的 DET 曲线看出，它的性能要远远好于传统的 GMM 模型。支持向量机(SVM)作为一种区分性辨别模型，被广泛地应用在生物识别领域。与 GMM 这种概率统计模型相结合可以提高系统的识别性能，从实验的DET 曲线可以看出，单纯的GMM 和SVM 两者技术的结合，相对于GMM-UBM 系统来说虽然没有很高的提高，但是得出的最小 DFC 要优于 GMM-UBM 系统。最后在SVM 模型中引入了基于 GMM 超向量，系统的性能得到了很好的提高。

6.2 工作展望

为了获得更好的效果和实用价值，还有许多工作需要进一步进行研究：

（1） 参数的选择和鲁棒性算法：例如可以对 Mel 倒谱系数进行小波变化；对于参数级融合如何取得更好效果；也可以考虑在参数上去除信道因子等等。

（2） 高斯混合模型：GMM 中通常的使用都对角化的高斯方差，隐含的假设条件是参数各维之间不相关，这在有限的数据集下的估计比较可靠。如何改进高斯混合模型并且不导致维数灾难是个值得研究的问题。

（3） 如何更好地结合 SVM 获得说话人确认的效果，还可以尝试引入其他机器学习的算法，增强说话人确认系统的性能。

（4）对于较长的时间刻度上的频谱变化或者基音的包络、变化轨迹(韵律)，人说话的速率(语速)以及更高层次的语音文本内容等等如何建模，更好地结合到传统的说话人模型中。

（5） 如何将这些说话人识别技术从实验室研究转向实际的应用，从服务器或个人电脑转向嵌入式系统中的实际应用，从需要大量计算的非实时性算法融入到快速的实时性高的系统中。

致　谢

在这次写论文的过程中，程建政老师给予了我悉心的指导。在此表示衷心的感谢。程老师对我毕业论文有着严格的要求，并且鼓励我进行大胆创新，这些都使我受益良多。特别是对格式和语言等一些小细节都很注意，很多东西由于我的粗心忽略了，多亏老师及时的指出来，经过多遍修改，才使论文逐渐完善。这种严谨的作风对我以后的工作和学习都很有帮助。

确实在写论文的过程中碰到了不少问题，从找资料到整理，到最后的论文，我经历了由不懂到懂的过程。老师总是牺牲自己的休息时间来辅导我们，给予悉心指导，使我逐步掌握了自我学习自我研究的办法。本文能顺利完成，凝结着恩师太多的心血和汗水。在此向我的导师表示深深的敬意和衷心的感谢。电子与电气工程学院的各位老师，在百忙当中对论文结构、形式、语言文字等方面给予了很多指导、并提出了很多宝贵意见，在此一并表示感谢。由于本人学识水平有限，文章中错误和疏漏之处在所难免，恳请各位老师提出宝贵意见，并给予批评指正，本人将不胜感激。

参考文献

[1] 赵力．语音信号处理[M]．北京：机械工业出版社，2009．5

[2] 张雪英．数字语音处理及 matlab 仿真[M]．北京：电子工业出版社，2010．7

[3] 吴朝晖，杨莹春．说话人识别模型与方法[M]．北京：清华大学出版社，2009．3

[4] The NIST Year 2010 Speaker Recognition Evaluation Plan

[5] 王炳锡．实用语音识别基础[M]．北京：国防工业出版社，2005．1

[6] 徐波，语音识别技术与应用的发现趋势[J]．中国计算机学会通讯，2008，2：54-57

[7] Joseph P. Campbell, JR. Speaker Recognition: A Tutorial, IEEE Proceedings of The IEEE[J], September, 1997, vol. 85, no. 9: 1437-1462

[8] J Campbell, D A Reynolds. Corpora for the Evaluation of Speaker Recognition[J]. Systems Proceedings of IEEE ICASSP , 1999, Vol.2: 829-832

[9] 韩纪庆，张磊，郑铁然．语音信号处理[M]．北京：清华大学出版社，2004

[10] Berouti M, Schwartz R and Makhoul J. Enhancement of speech corrupted by acoustic noise[J]. ICASSP, 1979: 208-211

[11] A. Martin, T.K.G. Doddington, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance[J] . Proceedings of EuroSpeech,1997, volume 4:1895-1898

[12] 刘保柱等．matlab 7.0 从入门到精通[M]．北京：人民邮电出版社，2010，5

[13] 章万锋．基于 PCA 与 LDA 的说话人识别研究[J].

[14] 张彩红，洪青阳，陈燕．基于 GMM-UBM 的说话人确认系统的研究[J]．心智与计算， 2007，1(4)：420-425

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

[16] 吕成国，王承发，李俊庆等．RASTA-PLP 技术与谱减相结合的去噪方法[J]．自动化学报，2000，9，26(5)

[17] 李霄寒，基于概率统计模型的说话人确认的研究[博士学位论文]．中国科学技术大学，2003

[18] 宫晓梅，王怀阳．噪声环境下 MFCC 特征提取[J]． 模式识别，2007

[19] L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition[J]. Prentice Hall Press,1993.

[20] 李杰，刘贺平．高斯序列核支持向量机用于说话人识别[J]．2010，46(18)： 183-185

[21] Reynolds D A. Experimental evaluation of features for robust speaker identification [J].

IEEE Transactions on Speech and Audio Processing . October 1994, 2(4):639–644.

[22] 卜奎吴，复杂环境下的说话人识别[J]．福建电脑，2010，(5)

[23] 芮贤义，俞一彪. 噪声环境下说话人识别的组合特征提取方法[J]. 信号处理，2006，10，5：673-677

[24] 叶寒生，陶进绪，张东文等．噪声环境下基于特征信息融合的说话人识别[J]．计算机仿真，2009，3，26（3）：325-328

[25] 陈迪，龚卫国，李波．噪声鲁棒性说话人识别语音高频加权 MFCC 提取[J]．仪器仪表学报，2008，3，29（3）：668-672

[26] S. Furui. Cepstral analysis technique for automatic speaker verification[J]. IEEE Trans.Acoust. Speech Signal Processing, 1981. 29(2):254-272.

[27] Boll S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech and Signal Processing. 1979, 27:113-120

[28] 邓菁，电话信道下多说话人识别研究[博士学位论文]．北京：清华大学计算机科学与技术系，2007

[29] J. Pelecanos and S. Sridharan. Feature Warping for Robust Speaker Verification[J]. Proc. ISCA Workshop on Speaker Recognition, 2001

[30] H. Hermansky and N. Morgan. Rasta Processing of Speech[J], IEEE Trans. on Speech and Audio Processing, Oct. 1994, vol. 2, no. 4:578-589

[31] P. Kenny, P. Dumouchel. Experiments in speaker verification using factor analysis likelihood ratios[J]. Proc. Odyssey04, 2004:219–226.

[32] Kocsor A, Toth L, Kuba A, etal. A comparative study of several feature transformation and learning methods for phoneme classification[J]. International Journal of Speech Technology, 2000. 3(3): 263-276

[33] Gish H,Ng K. Parametric trajectory models for speech recognition[J]. Proc of the Int'l Conf on Spoken Language Processing,1996,Vol.1:466-469

[34] Gales M J F and Young S J. Robust continuous speech recognition using parallel model combination[J]. IEEE Transactions on Speech and Audio Processing, 1996,4(5): 352-359

[35] Renevey P and Drygajlo A. Statistical estimation of unreliable features for robust speech recognition[J]. Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP ,2000, Istambul, Turkey, 2000:1731-1734

[36] Carlson B A and Clements M A. A projection-based likelihood measure for speech recognition in noise[J]. IEEE Transactions on Speech and Audio Processing

1994,2:97-102

[37] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel. Speaker and Session Variability in GMM-Based Speaker Verification[J]. IEEE Transactions on Audio, Speech and Language Processing, May 2007, Volume: 15, Issue: 4:1448-1460.

[38] Vapnik V. N. The Nature of Statistical Learning Theory[M]. NY: Springer-Verlag, 1995

[39] Vapnik. V. N. and Chervonenkis. A.Y. Theory of Pattern Recognition[M]. Nauka, Moscow,1974

[40] S Fine,J Navretil,R AGopinath Hybrid GMM/SVM Approach to Speaker Identification[J]. Proc.ICASSP,2001

[41] D Xin,Z H Wu.Speaker Recognition Using Continuous Density Support Vector Machines[J]. IEEE Electronics Letters,2001,37(17):1009-1011

[42] A. Ganapathiraju , J. Hamaker , J. Picone Ganapathiraju,Hybrid SVM/HMM Architectures for Speech Recognition[J]. ICSLP 2000:504-507

[43] Tommi Jaakkola , Mark Diekhans , David Haussler A Discriminative Framework for Detecting Remote Protein Homologies.journal of computational Biology,1998

[44] V Vapnik.Statistical Learning Theory[M]. New York:Wiley 1998

[45] W. M. Campbell, D. E. Sturim, D. A. Reynolds Support Vector Machines using GMM Supervectors for Speaker Verification[J].

[46] 黄亚娟，郑林．一种新的抗噪语音特征的提取[J]．微电子学与计算机，2011，10，25(10)：215-216

[47] 李姮，胡维平．基于 GMM 的说话人识别[J]．广西物理，2011，32(1)：21-24

[48] 张彩红，洪青阳，陈燕．基于 GMM-UBM 的说话人确认系统的研究[J]．心智与计算， 2007，1(4)：420-425

[49]陈雁翔，戴蓓倩，周曦等．基于对话语音的与文本无关的说话人确认系统的研究[J]．中文信息学报，2004，18(2)：36-42

[50]武妍，金明曦，王洪波．基于 KL-小波包分析的文本无关的说话人识别[J]．计算机工程与应用，2005，4：26-18

附　录

本人在攻读硕士研究生期间发表了两篇论文。在电子设计工程 2011 上发表了《周期性信号中随机噪声的延时消除技术研究》，在数字技术与应用 2012 发表了《小波分析在语音信号处理中的应用》。