分类号_____ 密级_____

UDC_____

# 学 位 论 文

## 基于GMM的声纹识别技术的研究

刘士

指导教师姓名 李绍荣 研究员

电子科技大学 成都

申请学位级别 硕士 专业名称 电路与系统

论文提交日期 2012.04 论文答辩日期 2012.05

学位授予单位和日期 电子科技大学

答辩委员会主席_____

评阅人_____

年 月 日

注 1 注明《国际十进分类法 UDC》的类号

## 独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

签名：_____ 日期：　　年　　月　　日

## 关于论文使用授权的说明

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技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，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

签名：_____　　导师签名：_____

日期：　　年　　月

# 摘　要

声纹识别又称为说话人识别，都是根据人生物特性来判断人的身份。声音作为人最自然的交流手段，以其无法比拟的优势被广泛应用到身份识别中去。本文的工作是研究基于高斯混合模型的说话人识别技术，并对特征参数的选取和识别算法做了一定的改进，以便获得较高的识别率。根据说话人识别的几个阶段，详细阐述了说话人识别技术的特征提取，模型建立等环节。

声纹识别的建模有很多种技术，目前高斯混合模型以其建模简单、性能好、与文本无关等特性是使用最多的建模方法之一。本文介绍了高斯模型的建立、参数估计以及识别方法。在识别阶段根据语音帧中的某些特定不好的语音帧会影响系统的识别率的情况，给出了一种基于帧投票的判决方案。由于高斯混合模型在说话人很多的时候计算量较大，本文使用 VQ 方法来对高斯混合模型分成男声和女声两个部分，并使用动态时间规整算法来计算各个基音之间的距离来减少模型的对比次数，从而减少了识别时间。

目前，大部分的声纹识别模型都是基于 MFCC 的混合高斯模型，MFCC 包含语音频率结构的时间变化信息，相对稳定，但不同的声纹之间容易相互模仿，本论文针对 MFCC 的易模仿性，增加了另一种特征参数，基音周期，基音周期包含了语音频率结构信息，虽然会受到说话人健康状况的影响，但不容易模仿，将二者结合用于声纹识别。针对 MFCC 特征参数会损失人的部分声学特性的情况，将动态 MFCC 系数加入到特征向量中，又由于加入后会使得特征向量变得复杂，根据他们对身份识别率的贡献给出了一种加权的 MFCC。

在文章的最后部分进行了实验验证。验证了特征参数、高斯混合模型阶数、加权的 MFCC 等对识别率的影响。实验发现，MFCC 的识别率高于 LPCC 识别率，MFCC 结合动态 MFCC 后的系统识别率有着明显的提高，加权的 MFCC 识别率高于原 MFCC 识别率并且和结合动态 MFCC 的识别率相近，这说明加权的 MFCC 在提高了识别率的基础上又减少了计算的复杂度，最后分析了基音周期的作用与影响。

**关键词：**声纹识别 ，说话人识别，MFCC，高斯混合模型，基音周期

# ABSTRACT

Speaker recognition technology, also known as the voice print recognition, is based on human biological characteristics to determine the identity of the person. Sound as the most natural means of communication, with its incomparable advantages was widely applied to identification.

For speaker recognition modeling there are a variety of techniques, Gaussian Mixture Model with its simply, good performance and text-independent feature is one of the most frequently used method of modeling. This thesis describes the Gaussian model, parameter estimation and recognition method. For speech frames in certain voice frame will affect the system recognition rate in the recognition phase,we give a voting based method.Using gaussian mixture model in the speaker recognition,when the speaker's number is large then there need amount of calculation.We combine the VQ method with Gaussian mixture model,the models are divided into two main parts which are Male and Female parts, and we use dynamic time algorithm to calculate the distance between each pitch,then reduce the recognition time.

At present,most work of speaker recognition technology study is based on the Gaussian mixture model. In order to obtain a higher recognition rate we choose better sound charcteristic parameters of the speaker and recognition algorithms. Speaker recognition elaborates the characteristics of speaker recognition technology, extraction, modeling and other sectors. Recently, most speaker recognition method are using MFCC and based GMM model.Another feature parameter of voice speech, pitch, is added in this paper against imitative of MFCC.Adding Dynamic MFCC coefficients to the feature vector will make the feature vector becomes complex, to shorten the time of speaker recognition,we given a weighted MFCC based on their contribution to the identification rate.

Experimental section is in the last part of this thesis, verify that the characteristic parameters of the Gaussian mixture model order, weighted MFCC, the recognition rate and the experimental results analysis.The experiment results show that the MFCC have a better performance than LPCC.When MFCC combine with Dynamic MFCC,the

recongnition rate was obviously increased.The Weighted MFCC raises the recognition rate and at the same time reduce the complexity of the calculation.We analysis pitch's function and its effect on recongnition rate at last.

**Keywords:** Speaker recognition, Voice print recognition,MFCC,GMM,Pitch

# 目 录

# 第一章 绪论

## 1.1 课题研究背景

语音处理技术在最近 30 年来已经得到了迅猛的发展，特别是在语音传输和数字语音存储方面的发展给人类的生活带来了极大的便利。语音识别因为其快速性和便捷性已经得到了广泛应用，如手机语音输入法。原来的许多人工服务也逐渐的被语音服务代替。在不久的将来估计语音认证将会被广泛应用到生活中的各个领域。传统的身份确认系统都是利用个人所知道的信息来作为身份确认的依据，如使用密码，或者使用个人所拥有的物品来判断，比如使用身份证等。但是传统的方法正面临着巨大的考验，主要原因是个人的账号越来越多，密码容易被遗忘等许多问题，相应的，利用人类生物学特性进行身份认证被越来越多的运用到人们的日常生活中去。

常见的生物特征有指纹、人脸、视网膜和声音等，在所有的生物特征中，由于每个人的声音特性和说话习惯都不一样，语音作为人类交流最自然的特征之一并且容易产生、获取等特性所以非常适合作为身份识别的工具，由于声音和指纹一样对每个人来说都不相同，相对于密码来说不用去记忆，所以研究声纹识别将可以给人们的日常生活带来极大的便利。

## 1.2 声纹识别概述

声纹识别又称为说话人识别（本文统称说话人识别），在应用上有着不同的分类大致可以分为两种：即说话人辨认（speaker identification）和说话人确认（speaker verification）[1]。

说话人辨认：从一堆已知的说话人中，根据说话人的语音选择最像的一个，这种情况是一种多选一的问题。说话人辨认进一步可以分为两类，一类为开放集（open set）的说话人辨认，另一类是闭集（closed set）说话人辨认，前者是说话人的范围没有限制，可以为任意的说话人，后者的范围限定为已知的一群人[2]。

说话人确认：根据说话人的语音和他所宣称的身份来验证是否该说话人是其所宣称的身份，这种情况是一个二选一的问题。

说话人识别根据在验证阶段的说话内容是否是识别系统所规定的，又可以分为与文本相关的（text-dependent）和与文本无关的（text-independent）[3]，前者是识别系统规定说话人所说的内容为已知，也就是说话人的语句是固定的或者是由系统提示的。后者则不限制说话人所说的内容。与文本相关的缺点是冒充者可以拿着真正说话人的录音来进行非法的识别。为了加强系统的安全性很多时候会将多种认证方法相结合，如在关键的场合加入密码，指纹等识别手段。

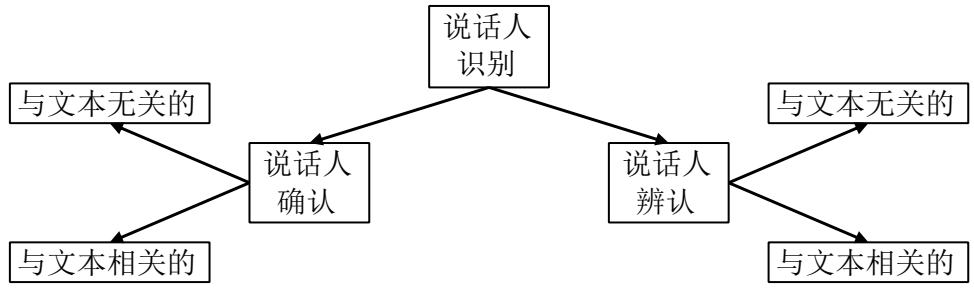
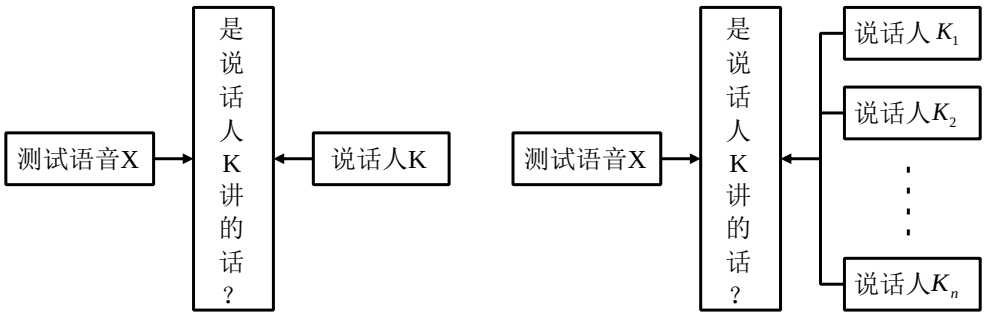



图 1-1 说话人识别分类




(a) 说话人确认示意图　　(b) 说话人辨认示意图

图 1-2 说话人确认和说话人辨认区别

## 1.3 说话人识别的应用

随着信号处理技术的进步和硬件水平的提高，说话人识别在金融领域、司法鉴定、军事安全甚至医疗领域都得到了广泛应用。例如，在金融领域，可以利用电话认证来登陆网上银行；在信息服务领域，可在安全领域，声纹可以作为出入机密场所的凭证；在公安司法领域，长期研究表明声纹可以作为罪犯嫌疑人身份鉴定的辅助手段；在军事领域，声纹可以用来鉴别不同的指挥员和作战信息；在医学应用中，声音可以用于某些相关疾病的诊断等等。因此，对说话人特征提取及说话人识别技术的深入研究有着重要的现实意义。

## 1.4 说话人识别的发展历史和现状

语音处理系统从模拟系统的模拟带通滤波器对信号频率的分析开始。自从有了数字信号处理方面的理论支持，如快速傅里叶变换等理论，以及硬件处理速度的提升，使得其得到迅猛发展。1976 年德州仪器（Texas Instruments）制作了第一个说话人识别的原型[4]。后来，特别是 NIST（National Institute of Standard and Technology）在语音处理方面做出了极大的贡献，直到今天 NIST 的说话人识别评估系统依然是语音识别极其方便的评估体系。

说话人识别的前进归功于特征提取和建模两种技术的同时发展，早期的与文本相关的说话人识别使用动态时间弯曲（DTW）和模板匹配技术。早期的特征向量提取技术包括：基音检测，线性预测，倒谱分析，以及线性预测能量误差等。

最近对说话人识别的研究主要集中在与文本无关的说话人识别方面。特征提取技术主要基于短时语音帧分析，语音信号被设定为准平稳的，一般情况下语音的帧长为 8-30ms，采样频率一般为 8kHz-16kHz。倒谱分析和梅尔倒谱分析（MFCC）是说话人识别中最常用的短时分析方法，线性预测（LP）并不常用，但是很多时候常和 MFCC 结合来使用。

说话人识别中的第二个关键问题是建模，常用的建模方法包括高斯混合模型（GMM）[5]，隐形马尔科夫模型（HMM）[6]，支持向量机模型（SVM）[7]，矢量量化模型（VQ）[8]，和人工神经元网络（ANN）[5]。

隐形马尔科夫模型常被用来做为与文本相关的说话人确认，而高斯混合模型，支持向量机，矢量量化主要做为与文本无关的说话人识别。其中高斯混合模型被认为是现在最优秀的建模方法[5]。高斯混合模型是一种高斯概率密度函数（PDF）加权和的向量集。常被看做是单状态连续隐形马尔科夫模型，或者看作为“软”矢量化模型[10]。

支持向量机技术在过去的十年中常做为说话人识别技术的主流，但是随着高斯混合模型技术的快速发展，已经被逐渐取代。有文献[9]指出支持向量机和高斯混合模型方法的结合比单独的高斯混合模型效果要好的多。

各种形式的矢量量化方法被运用在了说话人识别的分类中，最常见的方法是用 VQ 方法对每个说话人的语音数据建立一个码本（codebook）[11]。在说话人识别中，基于 VQ 的建模方法的识别率低于 GMM 模型。GMM 和 VQ 在技术上联系非常紧密，利用他们的相似性将二者结合对于说话人识别来说识别率要高于单独的 GMM 模型[12]。

人工神经元方法的大量的架构和多种形式被用在了说话人识别任务中[13]。多种 ANN 方法包括多层感知（MLP）网络，RBF 网络，时延神经网络（TDNN）等。文献[14]提出 RBF 网络比 MLP 网络效果更好。特别是在实验环境不完善的情况下 RBF 网络的表现更好，也就是说 RBF 的训练量要小于 MLP 网络。

文献指出[15]有些神经元网络的识别效果可以和 GMM 媲美，但是由于神经元网络和高斯混合模型在结构上的巨大差异，很难得出一个统一的结论哪个方法更好。从以上的比较可以看出高斯混合模型提供了一种非常好的说话人识别建模方法，所以最近的研究大多数集中在对经典的 GMM 算法的改进上[16]。更多的细节在第四章中说明。

现在运用最多的技术是使用 MFCC 特征向量和基于高斯混合模型的识别，这两项技术到目前仍然是最先进的技术。所以本论文所做的工作主要都是围绕着怎样改进 MFCC 参数和怎样更好的建立模型，以达到更好的识别效果。

从 863 计划开始执行以后，我国的语音识别方面得到了快速的发展，已接近国外先进水平，主要由中科院主持开发。不过大多数研究都集中在语音识别方向，比较成熟的有安徽的科大讯飞开发的讯飞系列产品，识别率已经相当高，现如今已经作为商用。

最近的说话人识别的主要研究主要集中于两个方向[17]：即“与本文无关的”和基于特定群体的嵌入式方向语音识别。另外一个方向则是高级别的说话人识别技术（即同时关注说话人所说的内容）。

## 1.5 说话人识别的系统架构

说话人识别分为说话人模型训练阶段和说话人识别阶段。根据说话人识别的分类它们在识别阶段稍有不同。图 1-3 为说话人模型的训练，该过程主要是对说话人的录音数据中进行特征参数提取，然后对这些特征信息进行模型训练得到说话人的语音模型。图 1-4 和 1-5 为说话人辨认和说话人确认的识别过程示意图。

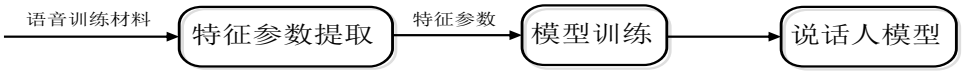



图 1-3 说话人模型的训练

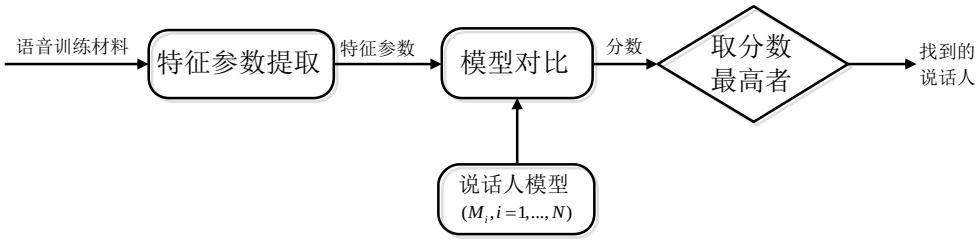



图 1-4 说话人辨认过程

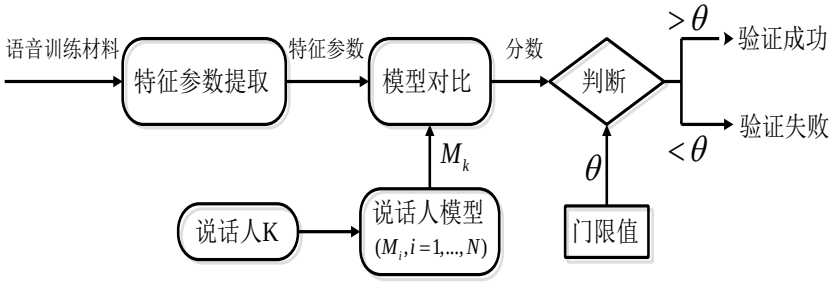



图 1-5 说话人确认过程

## 1.6 说话人识别中的语音特征

人类根据对每个人发声的特征不同来区分每个说话人。通常说话人识别中用到的语音特征分以下几类：低级别（low-level），中级别（middle-level）和高级别（high-level）信息。低级别特征如口音，个人说话习惯等，包含频谱。这些信息可用通过对信号的短时分帧来提取，他们对于大脑对语音的感知来说是次要的。中级别包含说话的节奏，声调等。高级别特征则用到了更长时间的语音信号帧，即是我们所熟知的词、短语这些声音中的特征，这些是人类大脑能够识别出的语言的关键。

低级别的特征被更多的用来作为说话人识别，因为他更容易从语音信号中提取出来，同时也因为在说话人识别中这些信息就已经足够了。为了理解如何捕获这些低级别特征，就需要了解人类是如何发声的。对于语音识别和说话人识别来说，低级别的信息主要为倒谱特征信息。只所以高级别的语音特征量更难提取是因为从一段语音信号中提取出单词或者短语需要建立一个大的词典库，一般的词典库可能会不全面。近年来由于语音识别技术已经足够快足够精确，所以高级别的语音特征提取技术也被给予了很大的关注。

根据文献[4]理想的说话人特征参数应该具有特点：（1）特征向量不应随着环境和信道的变化而产生较大的影响。（2）说话人的身体状况不应对识别结果造成太

大的影响。（3）提取的特征向量应该易于计算。（4）特征信息应该具有较强的反冒充者能力。

对于每个说话人来说其特征参数分类主要有两种：一种是人类固有的生理特性如基音和共振峰等。另一种是动态特性，这类常用的是倒谱以及其差分和基音。常用的倒谱系数有线性预测倒谱系数（LPCC）和梅尔频率倒谱系数（MFCC）[18]。本文对两种特征参数的提取都进行了探讨。

## 1.7 说话人识别中的建模

说话人识别建模的目的是用很少的参数和计算条件允许的情况下精确有效地区别开不同的说话人，即使他们所说的内容相同。目前建模方法大致可以分为两类：非参数（non-parametric）和参数（parametric）模型[19]。

### 1.7.1 非参数模型

非参数模型在参考库中的数据和要测试的语音数据相匹配的情况下效果比较好，如在“与文本相关的”说话人识别中，常常用简单的数字做为密码来建模，在测试时只需要对比说话人的语音特征是否和模板数据相匹配即可。另一种非参数模型是基于 Parzen 窗口的最大似然估计建模方法，每个说话人的语音数据概率密度函数都用 Parzen 窗来表征。一组语音参数可以对应于一种语音信号频谱，将这样的参数组看成一个矢量，不仅在数学上非常自然，而且在主观上有明确的物理意义，这就是语音信号的矢量表示（VQ），矢量量化是由标量量化推广和发展而来的一种信源编码技术[20]。矢量量化是一种极其重要的数字处理方法，已广泛应用于图像压缩、语音压缩等领域[18]。在 VQ 中说话人的语音信号被划分为多维向量区间，每个区域的中心部分表征了该区域并被存储起来，被测试向量是与存储的向量数据对比，找出最接近的一个。总体上说非参数模型应用不广泛。

### 1.7.2 参数模型

参数模型由于其健壮性被广泛的应用到了说话人识别中。该模型用几种参数来表征语音数据的平滑分布。其中用的最多的是高斯混合模型（GMM），大多数的说话人识别都用到高斯模型或者与其他模型相结合的高斯变形模型。我们将在后面的章节中详细论述。

改进的高斯混合模型（Adapted GMM）被广泛应用于基于高斯模型的说话人识别系统中。由于在很多系统中已经建模的语音数据并不多，所以使用 GMM 建模将不能提供足够的可靠数据。一种解决方案是由某个组织来建立一个通用的语音模型库，这样当我们建模时只需要很少的数据就可以完成。

隐马尔科夫（Hidden Markov Model，HMM）模型是语音识别中应用最广泛的模型，同时也被运用到了说话人识别当中。隐马尔科夫模型是一种用参数表示的，用于描述随机过程统计特性的概率模型，它是由马尔科夫链演变来的[16]。由于语音数据可以看做一个随机过程，这样我们可以为说话人建立一个模型，从而可以得出状态转移概率矩阵。在识别的时候通过计算状态转义过程中的最大概率来进行判决。隐马尔科夫模型可以节省计算时间，但是计算量大。

多层感知模型（Multi-layer perceptron ，MLP）是一种神经元网络技术，常被运用到语音处理中。MLP 的权重值通过回朔算法来计算，当输入为某些特定值时，输出可以达到很大值。说话人识别常采用 MPL 技术，理想情况下，在 MLP 中每一个说话人的特征向量输入对于该说话人都会有一个输出一个响应，而对于别的说话人都会有一个 0 状态输出响应。在测试阶段，所有的测试向量都要经过 MLP 网络，并将所有的输出叠加。选取叠加后的输出值最大的那个做为识别结果。当说话者数量增加时，该方法识别效果急剧下降。此外还有支持向量机模型，这里不再详细介绍。

## 1.8 说话人识别中的补偿技术

在过去的 20 年里，很多研究都是对于怎样降低录音信道的不匹配，因为这种不匹配，极大地影响了说话人识别系统的性能。通常都是通过补偿技术来解决。总的说来补偿技术可以分为以下几类类：基于特征的，基于模型的，以及基于计分评价的。在本节中我们将简单回顾下这几种技术，值得一提的是通常我们不单单使用一种，而是结合他们来提高系统性能。

### 1.8.1 非模板补偿技术

基于特征的补偿目标是提取与说话人无关的信息，如录音设备信息，说话的内容，以及信道的影响，同时还要很好的区别开不同的说话人。这里我们只介绍几种最常见的基于特征补偿技术：均化倒谱方法（CMS）以及 RASTA-PLP 方法[12]。

均化倒谱法使用一个滤波器来去掉梅尔倒谱中的信道数据，将提取后的 MFCC

参数减去平均倒谱向量将会消除因信道不同带来的影响。当加性噪声存在的时候，将会用到均化倒谱的改进方法来消除噪声。

RASTA 是一种调制频谱分析用来减少信道卷积噪声。通过减小调制信号中的低频部分，并且加大频谱中的动态部分来完成。和均化倒谱相似都是用过滤波来去除加性信道向量。经典的滤波器的传输函数如式 1-1 所示：

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

该传输函数将会带来相位失真，从而会导致人类听觉感知的时间掩蔽效应。因此在 RASTA 滤波后一般要进行相位调整。在实际应用中一般同时使用 CMS 和 RASTA 来进行说话人识别的补偿。

### 1.8.2 基于模板的补偿技术

基于模板的补偿技术通过改进说话人的概率分布模型以及研究各种信道的特性来完成。最常用的是说话人模型综合（SMS）以及并行模型组合方法（PMC）。SMS 研究在不同的信道条件下说话者参数模型是如何变化的，当没有注册的数据时就可以用这些参数来建模。建模使用 UBM（universal background model）来合成识别模型，该算法假设所有的说话人具有相同的转换模型，即使在不同的信道条件下，而实际情况可能会有所不同。

## 1.9 说话人识别的难点

（1）说话人的个性特征很难提取和分离 [13]

由于说话人的语音信号包含了说话内容和包含了说话人独有特征的向量，怎样将两者很好的分离开，对于不管是语音识别还是说话人识别来说都是一个值得研究的问题。

（2）环境的不同造成的说话人识别率的下降

由于实际的环境不同于实验室环境，往往干扰比较大，很多时候在实验室条件下表现比较好的识别方法在实际环境下表现往往并不理想。另外就是在特征提取的过程和测试过程的环境很难保证一致，如麦克风的不同，环境噪声的不同等。

（3）说话人个性特征的变化和语音样本的选择问题

每个说话人的固有特征可能随着时间的不同而变换，说话人的状态影响着该说话人的声学特征，如一个人在感冒和非感冒状态下的声音特征就可能不同，或

者生气或者高兴的两种状态声学特征可能也会产生变化。如何有效降低这些因素的不良影响，提高说话人特征参数的个人稳定性，是一个值得研究的难点问题。

（4）说话人训练时间的问题

一般的说话人识别都要经过一段较长时间的训练和测试才能很好的进行识别，这显然和实际需求有矛盾，如何找到一种训练和测试的方法是训练和测试过程简单也是摆在我们面前的一个问题。

（5）如何处理故意伪装的声音[4]

在法庭的说话人识别应用中，这个问题显得尤为重要，因为罪犯可能伪装自己的声音或模仿另一个人的声音。

## 1.10 本文结构安排

本文的组织安排如下：

第一章：简要介绍了说话人识别的分类，回顾了说话人识别发展历史和识别中所用到的方法，以及实际应用中所面临的问题。

第二章：介绍了语音信号处理所用到的基本原理和方法，以及语音的端点检测方法。

第三章：介绍了语音信号的特征向量提取过程，详细分析了基音检测方法和梅尔倒谱系数的提取过程以及其改进方法。

第四章：详细讨论了基于高斯混合模型的说话人识别系统的模型建立，说话人识别的判断依据，并对高斯混合模型进行了改进。

第五章：针对说话人识别中的各种方法进行了实验验证和分析。

第六章：总结展望。

# 第二章 语音信号的分析

语音是人类沟通交流的最重要的方法之一，语音处理与语言学、生理学、声学、电子、数学、计算机等密不可分。对语音信号进行处理，然后可以进行语音识别、语音编码（压缩解压）、语音增强（例如去掉背景噪音）等。一般来说语音信号的分析可以分为时域分析和频域分析两种方法，并由此衍生出清音、浊音，基音、泛音，平均能量、平均幅度、短时过零率等一系列语音参量。

## 2.1 语音信号的产生

人类的发声过程主要是由肺部的活动引起气流经过声门声道，引起声带振动的结果。当空气从肺部呼出时，气压带动声带。由于压力的不同从而声音的强度也有所不一样，声带的张力和质量决定了声音信号的频率特性。声带的每开启和闭合一次的时间称为基音（pitch）周期，其倒数称为基音频率[25]。声带的形状以及声门的气压差效应等决定了基音周期的大小。基音频率会随着年龄、性别、个体等情况的不同而变化，范围可达 80Hz-500Hz，所以说话人识别时常常将基音频率来作为区分不同说话人的重要参数。

当肺部发出的气流经历声带时引起共振而产生的声音称为浊音（voice），没有声带振动的声音称为清音（unvoice）[19]。浊音反应了声带的特性并且具有周期性，所以是语音处理中的主要用到的部分，而清音不是由声带振动发出，相当于随机噪声。不同的浊音波形是不同的。清音的语音信号具有随机噪声的特点。一般来说清音的幅度小于浊音的幅度[27]。

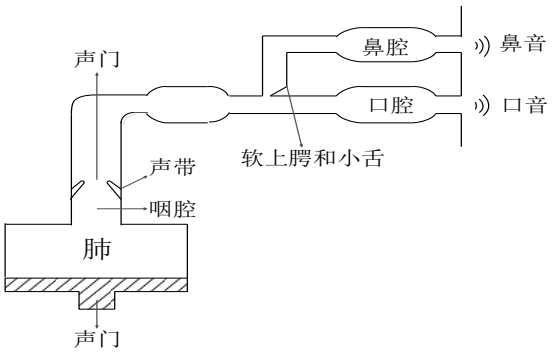



图 2-1 语音产生的机理图

在很长时间内，人们不了解声音听起来不同的原因，法国物理学家傅里叶发现声音不同的根本原因在于和弦的不同。和弦由一个声音的基音和倍频音组成。当声波通过声道后，就引起声道的共振。声道的谐振倍频频率称为共振峰频率或共振峰[28]。我们知道即使两个人说话内容相同，熟悉他们的人依然能够区别开来，这主要由音色来区分，而声道的共振峰特性决定了音色。

## 2.2 语音信号的数字模型

经过前人的大量研究，综合考虑声门激励、声道以及嘴唇辐射的影响得到下图所示的语音产生的离散系统模型[21]。

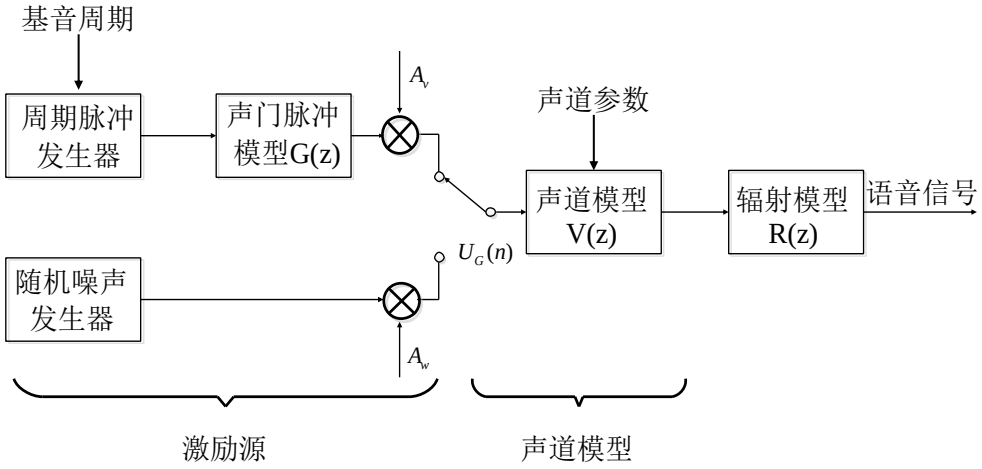



图 2-2 语音信号产生的离散时域模型

语音信号可以看做是激励信号$U_G(n)$经过一个线性系统$H(z)$而产生的输出[21]。其中声道模型$H(z)$为声道响应模型和嘴唇辐射模型的级联，对于浊音来说还以把声门脉冲的影响也加入传递函数中。总之$H(z)$可以简化成一个全极点的线性函数。对于不同的人来说$H(z)$很好的表征了每个人之间声学特征的不同。$H(z)$的表达式为：

$$H(z)=\frac{1}{A(z)}=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{2-1}$$

其中$a_i,(i=1,2,...,p)$为滤波器的系数，p 为全极点滤波器的阶数，p 的取值关系着模型的准确度，通过实验可以确定，p 取 8-12 就能很好的将人的声道特性表征出来。上述语音产生模型的基本思想是将激励与系统相分离，使语音信号解体来分

别进行描述，而不是直接研究信号波形本身的特性，这种思想是带动语音处理技术飞速发展的关键[16]。

## 2.3 语音信号的预处理

### 2.3.1 采样与量化

语音信号是随着时间变化的一维信号，频率可达到 10kHz。在进行处理之前一般都要经过采样和量化处理。采样就是将时域模拟信号进行等间隔的抽样。根据耐奎斯特采样定理可以知道，要保持信号完整性，采样频率要大于等于语音信号最高频率的两倍。但是由于人类的语音中有大量的冗余信息，即使丢弃这些信息对实际处理也不会产生影响。国际电报电话咨询委员会（CCITT）规定语音信号的标准采样频率为 8kHz，在减少数据量的情况下而不会造成过分的失真，电话的频带宽度只有 3-4kHz。在实际的语音处理过程中，一般采样频率为 8-10kHz[21]。

本文实验所用语音为在实验室通过麦克风和录音机软件录制得到，采样量化过程由计算机的声卡自动完成，采样频率为 8kHz，采样用 8 比特表示，采样过后的数据保存为波形文件。

### 2.3.2 语音信号的预加重

经过采样量化后的语音信号，一般都要经过预加重处理。由于语音信号的高频部分（800Hz 以上）按照 6dB/倍频程进行跌落[25]，为了提升语音信号的高频部分，所以进行预加重处理，这样信号的频谱将会更平坦。预加重的公式如下：

$$H(z)=1-\mu z^{-1} \tag{2-2}$$

其中系数 $\mu$ 的值在 0-1 之间，可以通过经验或者实验来确定，一般取值为 $0.9<\mu<1$。其幅频特性为：

$$|H(w)|=\sqrt{1+\mu^2-2\mu\cos w} \tag{2-3}$$

由公式 2-3 可以看出当信号为高频时，幅值 $|H(w)|$ 趋近于 $1+\mu$，低频时趋近于 $1-\mu$，所以通过预加重处理以后，信号的高频部分得到了提升。

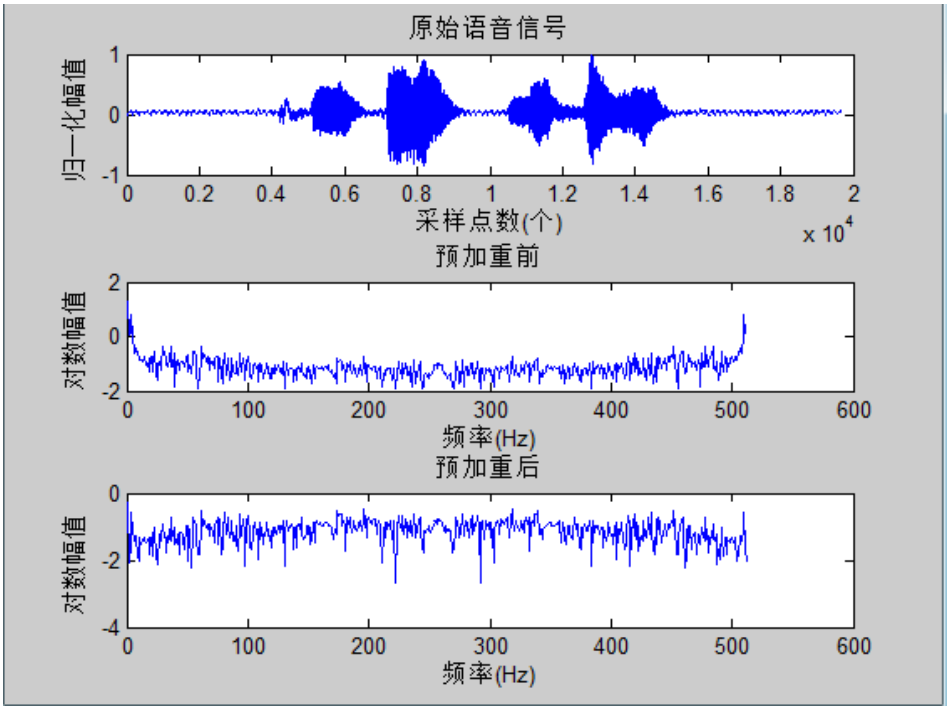



图 2-3 语音信号的预加重处理

图 2-3 是一语音信号经过的预加重处理后的示意图。从图中可以看出经过预加重处理后的语音信号的低频部分被抑制，高频部分得到了提升。

### 2.3.3 分帧后加窗

因为语音信号是短时平稳的，所以我们就可以将语音信号进行分帧处理。为了加强语音信号的频谱特性，可以将时域采样信号进行短时加窗处理（大概20-30ms），在该时段中随着时间的变化，语音信号变化缓慢。所以短时傅里叶语音分析常用一个滑动窗（语音帧）来进行。每一帧允许和前一帧数据有一定程度的重叠（50-90%之间）。当用 8kHz 的频率进行采样，窗口大小是 20ms，就可以有 10ms 的延迟（50%的重叠）。通常我们通过计算语音信号的频谱特性，并组合成一组向量，该向量叫做特征向量（feature vector）来表征语音信号的各个帧。

每一帧采样数据可以看做是一个线性时不变系统的的激励输出相应。因为线性时不变系统的激励相应是其冲击响应的卷积，所以语音处理就可以转化成卷积积分来处理。在实际应用用为了得到需要想要的频率响应，对信号的加窗函数也各种各样。矩形窗函数（N 点）通过式 2-4 给出:

$$W(n)=\begin{cases}1, 0\le n\le N-1\\ 0, \text{其他}\end{cases} \qquad (2\text{-}4)$$

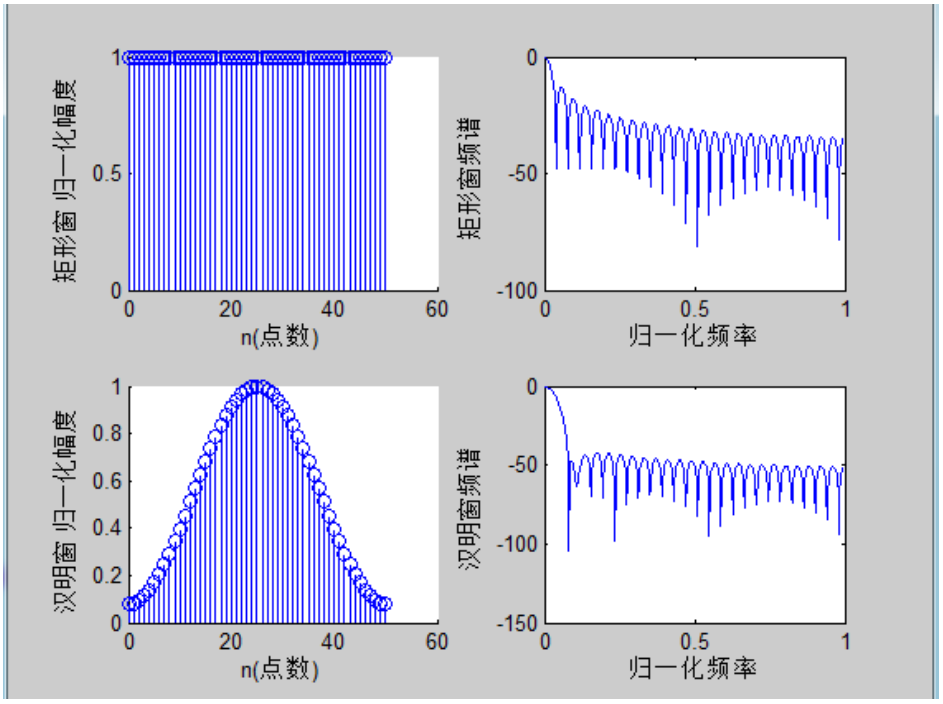



图 2-4 矩形窗和汉明窗频谱响应

其他形式的窗函数还有很多，如汉明窗函数。通常汉明窗是在语音处理中应用的最多的一类窗口函数，其定义为：

$$W(n)=\begin{cases}0.54-0.46\cos(\dfrac{2\pi n}{N-1}), 0\le n\le N-1\\ 0, \text{其他}\end{cases} \tag{2-5}$$

矩形窗，汉明窗及的时域和幅度特性如图 2-4 所示。由图 2-4 的频谱分析可以看出，矩形窗的主瓣宽度小，所以矩形窗的频谱分辨率比较高，但是同时矩形窗的旁瓣峰值较大，所以波形的细节容易丢失从而导致了频谱泄露[21]。相比较而言，汉明窗的主瓣宽度是矩形窗主瓣宽度的两倍，低通特性的平滑性会更好一点，所以可以很好的反映短时语音信号的频谱特性。根据窗口的特点汉明窗在说话人识别中应用比较多。

## 2.4 语音信号的时域分析

对语音信号最简单的分析就是以时间为自变量来分析，时域处理最常用的方法是短时能量分析，短时过零率，短时自相关函数以及短时平均幅度等。

## 2.4.1 短时能量和短时过零率

语音信号的能量随着时间而不断变化，一般情况下语音中的浊音的能量会比较大，语音的短时分析提供了一种区分清音和浊音的方法。语音信号$\{s(n)\}$的短时能量由公式 2-6 给出：

$$E_n = \sum_{m=-\infty}^{\infty}[s(m)w(n-m)]^2 = \sum_{m=-\infty}^{\infty}s^2(m)h(n-m) = s^2(n) * h(n) \quad (2\text{-}6)$$

短时能量代表声音的尺寸，可由声音信号的振幅来类比[27]。其中$h(n)=w^2(n)$，$E_n$表示第 n 点的信号加窗后的短时能量，w(n)为加窗函数。短时能量可以看做信号经过单位冲击响应为$h(n)$的一个线性滤波器后的输出，短时能量还可以用来判断一段语音中的有声和无声部分，从而可以用来做端点检测。

语音信号的幅值会随着时间的变化而变化，信号 s(n)的短时过零率定义为一段语音信号中波形与横轴相交的次数[12]，即采样信号的正负符号变化的次数。过零率可以用来大致的估计信号频谱特性和用来判断有话无话或者清音浊音。其计算公式如下：

$$Z_n = \frac{1}{2}\left\{\sum_{m=-\infty}^{\infty}|\mathrm{sgn}[s(n)]-\mathrm{sgn}[s(n-1)]|\,w(n-m)\right\} \quad (2\text{-}7)$$

其中 sgn( )是符号函数：

$$\mathrm{sgn}(x)=\begin{cases}1, & x\geq 0\\ -1, & x<0\end{cases} \quad (2\text{-}8)$$

由于实际环境的不同，可设定一个门限 T，此时过零率就变成了过正负门限的次数，于是公式 2-7 变成：

$$Z_n = \frac{1}{2}\{\sum_{m=-\infty}^{\infty}|\mathrm{sgn}[s(n)-T]-\mathrm{sgn}[s(n-1)-T]|+|\mathrm{sgn}[s(n)+T]-\mathrm{sgn}[s(n-1)+T]|\}w(n-m) \quad (2\text{-}9)$$

这样即使存在一些噪声，只要噪声的幅度不超过上下门限，就不会影响过零率的计算。

## 2.4.2 短时自相关函数

相关函数用来计算两个函数的相关程度，自相关函数用来研究函数本身的特性如同步与周期性。对于信号序列$\{s(n)\}$自相关函数定义如下：

$$R(k)=\sum_{m=-\infty}^{\infty}s(m)s(m+k) \tag{2-10}$$

通过对信号进行加窗处理然后在做自相关分析即可得到短时自相关函数：

$$\begin{aligned}R_n(k)&=\sum_{m=-\infty}^{\infty}s(m)w(n-m)s(m+k)w(n-(m+k))\\&=\sum_{m=n}^{n+N-k-1}s_w(m)s_w(m+k)\end{aligned} \tag{2-11}$$

其中 $s_w(m)$ 为经过加窗处理后的信号，该窗口在第 n 点加入。由自相关函数的性质可知 $R_n(0)$ 为最大值并且等于加窗后的信号能量。如果语音信号是一个浊音的周期信号，由自相关函数的性质可知，其短时自相关函数也是周期的，而清音信号接近随机噪声，所以没有明显的周期性，从而通过短时自相关函数可以确定一个浊音的基音周期。

## 2.4.3 语音信号的端点检测

端点检测的作用就是去掉语音中开始和结束的静音数据部分，从而达到更好的识别效果。本文利用短时能量谱和短时过零率检测语音端点，以短时能量为主，以短时过零率为辅，来去掉语音段中无声部分。

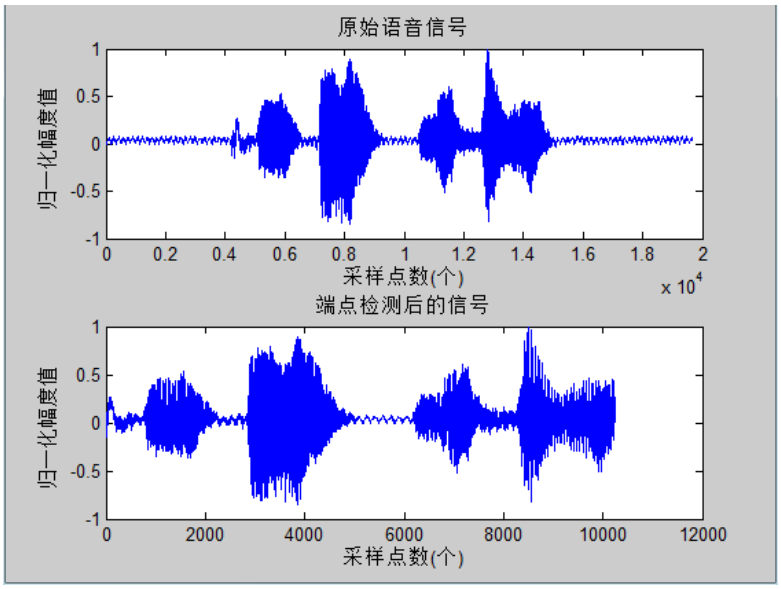



图 2-5 语音信号端点检测

观察图 2-5 可以发现语音信号的开始和结束部分幅度都很小，所以可以设定一个门限来检测语音信号的开始和结束。通常情况下清音的随机性比较大，并且幅

度比较小，所以会频繁穿过 0 点，即短时过零率比较高，而浊音不但能量比较大而且短时过零率比较低。

试验中，短时能量及短时过零率的阈值的选取凭实验经验选取，分别为最大值的五分之一，图 2-5 为一语音信号段和经过端点检测后的时域图。可以发现经过端点检测后幅度较小的部分（静音部分）被去除，只保留了有效的语音数据，这样处理后不仅减少了计算量，而且还提高了识别效率。

## 2.5 语音信号的频域分析

### 2.5.1 语音信号的短时傅里叶变换

由于傅立叶变换是分析信号的有力工具，同时域分析一样，频域的傅立叶分析也用到了短时分析技术。语音信号序列$\{s(n)\}$的短时傅里叶变换定义为：

$$S_n\left(e^{jw}\right)=\sum_{m=-\infty}^{\infty}s(m)\,w(n-m)e^{-jwm} \tag{2-12}$$

其中$w(n)$为窗函数，短时傅里叶变换可以映射为时间和频率的二维函数，也称为时频函数，值得一提的是短时傅立叶变换的平方称为短时功率谱。

### 2.5.2 小波变换在语音分析中的应用

小波变换在非平稳信号处理中有着傅里叶变换无法比拟的性质。小波变换在语音信号的处理中主要有以下几方面的应用：对听觉系统进行模拟，去随机噪声，对清音浊音的判断。本文主要用到了去随机噪声。

传统的方法是将语音信号通过一个滤波器进行滤波处理，去掉其中的噪声部分。但是由于语音信号是非平稳的，所以用到了小波变换的时频局部分析的特点。小波变换去噪的主要工作就是去掉噪声产生的小波谱分量部分。白噪声在小波变换下的特点和语音信号不同。假设一个信号$n(t)$为宽平稳白噪声，其方差为$\sigma^2$，$\psi(t)$是一个小波函数高斯白噪声的小波变换期望值为：

$$E\{|W_s n(t)|^2\}=\frac{\|\psi\|^2}{s}\sigma^2 \tag{2-13}$$

通过上式得知高斯噪声的能量随着小波变换的尺度加大而迅速减小，所以若增大 s 而迅速减小的是噪声，文献[21]中提到可以通过分析小波变换的模值极大值的方法进行去噪。

### 2.5.3 语音信号的同态解卷积

由前面的讨论可知当激励信号经过线性系统产生了语音信号。将卷积分量分开的过程称为解卷积。一般情况下解卷积的方法有两种，一是线性预测分析，将会在下章中做介绍，另一种是同态解卷积。同态分析也称为倒谱分析，经过同态分析后就可以得到语音信号的倒谱参数。由于语音信号并不是加性信号，而是两个信号的卷积，所以不能用线性系统来处理，同态处理就是将非线性问题转化成线性问题。假设输入信号是两个信号 e(n)和 v(n)的卷积，分别代表声门激励和声道相应序列。同态分析包括三步：

（1）Z 变换，将信号由卷积形式转化成了乘积形式，这样就可以得到的输入信号的频谱，变换公式为：

$$Z\left[s(n)\right]=S(z)=E(z)\times V(z) \tag{2-14}$$

（2）将乘积信号变为加性信号，由对数运算实现。

$$\log S(z)=\log E(z)+\log V(z)=\hat{E}(z)+\hat{V}(z)=\hat{S}(z) \tag{2-15}$$

（3）反 Z 变换，获得信号的倒谱（Cepstrum）。

$$Z^{-1}\left[\hat{S}(z)\right]=Z^{-1}\left[\hat{E}(z)+\hat{V}(z)\right]=\hat{e}(z)+\hat{v}(z)=\hat{s}(z) \tag{2-16}$$

## 2.6 语音信号的倒谱分析

根据声音产生的数学模型可以看出，语音信号是由声道冲击响应和声门激励信号的卷积，所以可以通过同态分析来进行解卷积，从而将声门激励和声道冲击响应分离出来。设语音信号 $s(n)=e(n)*h(n)$，其中 e(n)为声门激励信号，h(n)为声道冲击响应[29]。因为声道的特性决定了人的声音特征所以对语音信号的倒谱分析就变得十分有显示意义。将分析过程中的傅里叶变换换成 Z 变换，从而可以得到信号的复倒谱。倒谱的计算过程如图 2-6 所示，可以分成两部分。

（1）声门激励信号的倒谱

人类发出清音信号时，声门激励是白噪声，在发出浊音时，声门激励是以基音周期为周期的冲击序列。声门激励的倒谱也是一个周期冲击序列，并且周期长度和基音周期相同，并且倒谱的振幅随着时间的变化会逐渐减小，衰减速度比较快[25]。声门激励的浊音倒谱只有在基音周期的整数倍上不为 0，根据这一特点可以进行清音浊音的判断。

（2）声道冲击响应的倒谱

一般情况下使用全极点模型函数来描述声道特性，声道响应的倒谱维数较低，同时维数较少的倒谱向量就可以描述语音信号的声道特性，由于这个原因，对声道特性进行研究从而提出了线性预测倒谱系数。

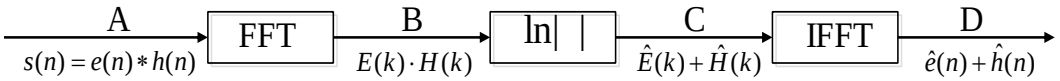



图 2-6 倒谱的计算过程

## 2.7 语音信号的特征评价标准

由于目前的语音信号特征提取技术的发展，各种特征向量都用来表征语音信号，但是没有一种理论来说明某一种特性向量参数一定比另一种更好。一般的方法都是通过实验的结果来验证识别率的高低，从而评价该种特征的好坏。但是实验的方法很难做到外部条件相同，容易受条件的制约。通常用到的方法为 fish 比和可分性测度。

# 第三章 特征参数提取方案与设计

在第二章中介绍了语音信号的处理方法，对说话人识别来说一个很重要的步骤就是提取语音信号中特有的可以用来表征说话人身份的特征向量，如基音周期，共振峰等。本章主要介绍被广泛使用的线性预测倒谱系数，梅尔倒谱系数，基音周期以及提取步骤，并对应用较多的梅尔倒谱系数给出了改进的方法。

## 3.1 基音周期估计

基音是指发浊音时声带震动的周期性，而基音周期是指声带震动频率的倒数[21]。语音信号的一个重要参数就是基音周期，因为基音周期并不具有严格的周期性，所以要用短时平均法来估计，也就是基音检测（pitch detection）。

### 3.1.1 基音周期的检测方法

（1）波形估计法：通过利用语音信号的波形来估计基音周期，分析出波形上的周期峰值[30]。

（2）相关处理法：前面已经介绍过浊音信号具有周期性，对于周期性信号序列，自相关函数定义为[30]：

$$R(k)=\lim_{N\to\infty}\frac{1}{2N+1}\sum_{m=-N}^{N}s(m)s(m+k) \qquad (3\text{-}1)$$

其中 k 为延迟点数。由自相关函数的性质可以知道，自相关函数反映了信号延迟一段时间以后和原信号的相似程度，周期信号的自相关函数也具有周期性，并且其周期和原信号相同[31]，有 $R(k)=R(k+Np)$。浊音是声带的周期性开启与闭合产生的，因而具有周期性，它的自相关函数的周期和浊音的周期相同，且在周期整数倍位置上出现峰值，第一个最大峰值到原点的距离可以看做为周期；清音信号不具周期性，这样就可以将两者区别开来[32]。

为了提高自相关方法检测准确性，一般要进行预处理。由于基音信息主要在语音信号的高频部分，共振峰信息主要在低频部分，声道响应中同时有共振峰周期和基音周期，这样就会造成峰值信息的混叠[33]。通过对低频部分的抑制，可以

改善自相关处理的性能。一种常用的中心消波处理方法为三电平中心消波。三电平中心削波的输出函数为[34]：

$$y(n)=C[x(n)]=\begin{cases}1, x(n)>CL\\ 0, others\\ -1, x(n)<-CL\end{cases} \quad (3\text{-}2)$$

即削波器的输出在 $x(n)>CL$ 时为 1，$x(n)<-CL$ 时为-1，其余值为 0，这样就可以滤除不重要的峰值与清音部分同时保留了具有明显周期的峰值。实验证明，这种方法具有简便，快速，准确和易于实现性。通常会用一个通带为 900Hz 的线性相位滤波器来去掉高次谐波，从而只保留第一共振峰以下的基波和谐波分量。本文用到了三电平削波的原理进行基音检测。

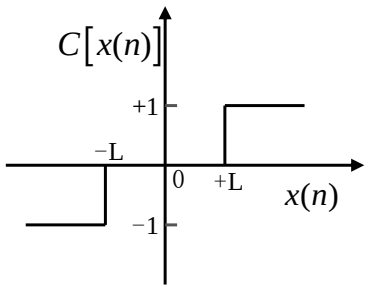



图 3-1 三电平消波函数

（3）倒谱法

对语音信号进行倒谱解卷积后可以得出激励序列的倒谱，他和基音的周期相同，所以可以利用倒谱进行基音检测。信号 s(n)的倒谱 e(n)由式 3-3 给出：

$$c(n)=z^{-1}[\ln|z(s(n))|] \quad (3\text{-}3)$$

其傅里叶变换形式为：

$$c(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln|S(e^{jw})|e^{jw}dw \quad (3\text{-}4)$$

语音的产生模型是一个由周期脉冲序列（浊音）或白噪声序列（清音）激励的线性滤波器，在一帧内（短时）滤波器可以近似认为是时不变的，因此，语音信号可以看做是滤波器与激励源冲激响应卷积的结果[35]。根据倒谱的性质可以将卷性的语音信号转变成为加性的语音信号，从而将滤波器与激励源分离开来。当语音信号为无噪声的纯净语音时用倒谱法进行基音检测效果比较好，但是当存在噪声时，由于噪声干扰了基因谐波的周期性，从而导致检测的结果不准确。本文

的基音周期检测用到了自相关处理方法。

### 3.1.2 基音周期的提取步骤

经过语音信号的预处理后（分帧加窗，滤波，预加重等步骤以后）基音检测的第一步为清音浊音判断，这通过短时能量来实现。短时能量的计算第二章中已经做过介绍。经过清浊音判断后取出语音信号中的浊音段，然后进行中心消波处理，中心消波的公式：

$$y(n)=C[x(n)]=\begin{cases}x(n)-L, x(n)>CL\\ 0 \qquad\quad , x(n)\le CL\\ x(n)+L, x(n)<-CL\end{cases} \tag{3-5}$$

图 3-2 所示为一段语音信号经过中心消波处理后的对比图。通过图可以看出经过中心消波处理后的一部分无效的清音语音数据被清零。

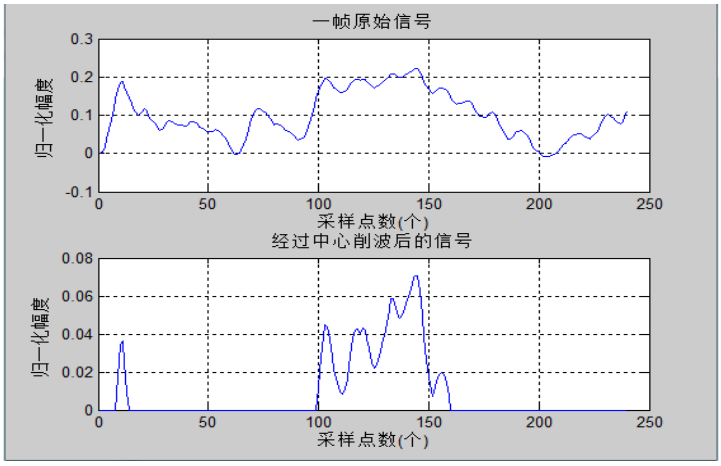



图 3-2 语音信号的中心消波处理

短时自相关函数在基音周期的整数倍点上为峰值，通过计算第一个峰值位置与原点之间的距离，得到估计的基音周期[36]。需要注意的是一般情况下在进行加窗运算的时候，窗口长度要大于两个基音周期长度。为了有效去除共振峰特性的干扰，还要经过带通滤波处理。为了减少运算量，在实际运算中用中心消波后的信号的互相关函数来代替自相关运算，对 $y(n)$ 进行三电平量化得到 $y'(n)$，即

$$y'(n)=C'(y(n))=\begin{cases}+1, y(n)>0\\ 0, y(n)=0\\ -1, y(n)<0\end{cases} \tag{3-6}$$

由于 $y'(n)$ 的只有三个可能取值，计算过后的互相关函数的周期性与 $y(n)$ 的自相关

序列周期接近。本论文中基音检测算法即为求 $y(n)$ 和 $y'(n)$ 互相关值：

$$R(k)=\sum_{n=1}^{N} y(n)\cdot y'(n+k) \qquad (3\text{-}7)$$

其中 $k=1,2,\ldots N$ /,如果 $R_{\max}<0.25R(0)$ 则认为本帧为清音，令其基音周期值 $P=0$，否则基因周期即为使 R(k)取最大值 $R_{max}$时的位置的 k 值[34]，即 $P=\arg\max R(k)$ 时。就是检得的基音周期估计值。图 3-3 为语音基音检测图。

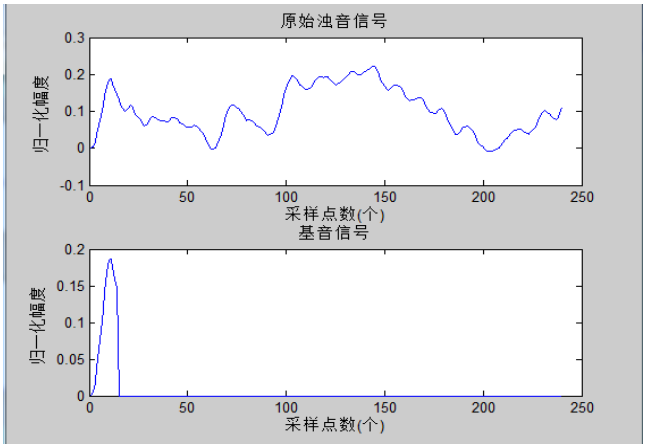



图 3-3 语音信号的基音检测

## 3.2 线性预测（LPC）以及其倒谱系数 LPCC

在基于参数模型的说话人识别系统中，通常会假设系统的传递函数为全极点模型，利用时域均方误差最小的原则来进行模型的参数估计，由于其计算量小，应用上比较灵活所以在语音处理中得到了广泛应用[37]。

### 3.2.1 线性预测的基本原理

线性预测的原理是利用语音信号 x(n)的前 p 个值来预测当前的值即：

$$\hat{x}(n)=\sum_{i=1}^{p} a_i x(n-i) \qquad (3\text{-}8)$$

图 3-4 为一个随机噪声和其线性预测估计得出的信号对比图。在实际的语音处理中，很多时候用 LPC 的倒谱来表征语音信号。

线性预测误差的定义由公式 3-9 给出 e(n):

$$e(n)=x(n)-\hat{x}(n)=x(n)-\sum_{i=1}^{p}a_i x(n-i) \tag{3-9}$$

由于语音的短时平稳特性，所以用短时预测均方误差：

$$E(n)=\sum_{n}e^2(n) \tag{3-10}$$

来作为预测结果水平的衡量。在均方误差最小的基础上可以得到最佳的线性预测系数$\{a_i\}$。图 3-4 是通过线性预测方法得出的信号和原信号的对比图。

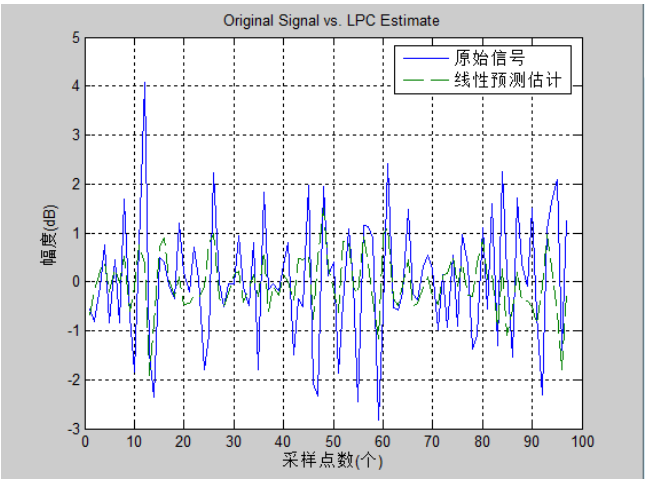



图 3-4 线性预测估计信号

## 3.2.2 LPCC 的提取

线性预测倒谱系数（Linear prediction Cepstrum Coefficient，LPCC）是基于语音信号为一种全极点模型（也称为自回归的）的假设，利用线性预测分析获得倒谱系数的一种倒谱特征[12]。由于人的声道特性只需要较低维数的 LPCC 参数就可以表征，所以 LPCC 被广泛应用到说话人识别中。

在第二章所述的语音信号的模型中，将语音信号的声道传递函数设为一个全极点模型如 3-11 式：

$$H(z)=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{3-11}$$

其中$a_i(i=1,2,...,p)$为线性预测系数，线性预测阶数的阶数为 p。若取$H(z)$对数形式，再将$z^{-1}$按照傅里叶级数展开后可以得到：

$$\ln H(z)=C(z)=\sum_{n=1}^{\infty}c_{LP}(n)z^{-n} \tag{3-12}$$

这样就可以得到语音信号的 LPC 倒谱系数$c_{LP}$。将式（3-11）带入式（3-12）可以得到：

$$\ln\left(\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}}\right)=\sum_{n=1}^{\infty}c_{LP}(n)z^{-n} \tag{3-13}$$

将式 3-13 两边$z^{-1}$对求导，再简化得：

$$\frac{\sum_{i=1}^{p}i a_i z^{-(i-1)}}{1-\sum_{i=1}^{p}a_i z^{-i}}=\sum_{n=1}^{\infty}n c_{LP}(n)z^{-(n-1)} \tag{3-14}$$

式 3-14 可写成如下形式

$$\sum_{i=1}^{p}i a_i z^{-(i-1)}=(1-\sum_{i=1}^{p}a_i z^{-i})\sum_{n=1}^{\infty}n c_{LP}(n)z^{-(n-1)} \tag{3-15}$$

在式 3-15 中，令方程两边$z^{-1}$各次幂的系数相等，得出线性预测系数$a_i(i=1,2,...,p)$与其倒谱系数$c_{LP}(n)$的方程：

$$\begin{cases} c_{LP}(1)=a_1 \\ c_{LP}(n)=\sum_{k=1}^{n-1}\frac{k}{n}a_{n-k}c_{LP}(k)+a_n,(1<n\le p) \\ c_{LP}(n)=\sum_{k=1}^{n-1}\frac{k}{n}a_{n-k}c_{LP}(k),(n>p) \end{cases} \tag{3-16}$$

线性预测倒谱计算比较简单，但是线性预测假设声道响应函数为全极点模型，由于实际中可能有零点响应，所以对比起梅尔倒谱特征来说效果要差些。

## 3.3 基于听觉特性的 Mel 频率

自 18 世纪以来，随着知觉心理学不断的发展，人们开始研究听觉感知系统，从而得到了人类听觉系统的许多特性。所有的研究集中在两个方面：可见的，包括人耳的构造，利用医学上的解剖来理解人耳各个部分的听觉特性，不可见的如人耳的掩蔽效应，人耳的听觉感知曲线等。由于人耳对语音信号有着很好的辨识率，所以有的学者认为人耳没有用到的信号是可以忽略的，在声学特征技术提取中，人们利用滤波器组来进行听觉的模拟，这就引出了梅尔刻度的倒谱系数

Mel 频标倒谱系数（Mel Frequency Cepstrum Coefficient，MFCC）不同于 LPCC

等特征参数，LPCC 是通过对人的发声机理的研究而得到的声学特征，MFCC 是对人的听觉系统研究得出声学特征[38]。MFCC 是一种以梅尔刻度倒谱参数，通过数字滤波器组的分析而得出的特征参数，对语音信号进行滤波器组分析是语音处理中比较常用的方法，人类听觉感知系统前端可以用滤波器组的来进行模拟[39]。

表 3-1 临界频带滤波器参数表

| 序号 | 中心频率（Hz) | 带宽（Hz) | 序号 | 中心频率（Hz) | 带宽（Hz) |
|---|---|---|---|---|---|
| 1 | 100 | 100 | 11 | 1149 | 160 |
| 2 | 200 | 100 | 12 | 1320 | 184 |
| 3 | 300 | 100 | 13 | 1516 | 211 |
| 4 | 400 | 100 | 14 | 1741 | 242 |
| 5 | 500 | 100 | 15 | 2000 | 278 |
| 6 | 600 | 100 | 16 | 2297 | 320 |
| 7 | 700 | 100 | 17 | 2639 | 367 |
| 8 | 800 | 100 | 18 | 3031 | 422 |
| 9 | 900 | 100 | 19 | 3482 | 484 |
| 10 | 100 | 100 | 20 | 4000 | 556 |

根据音调频率的对数和人耳对音调的感知强度成正比，提出了梅尔频率。梅尔频率与频率的关系为：

$$f_{Mel} = 2595\log_{10}(1+ f / 700) \qquad (3\text{-}17)$$

从公式可以看出在频率低于 700Hz 的时候梅尔频率和频率基本成线性关系，人耳感觉比较灵敏。而在大于 1kHz 高频部分两者呈成对数关系，对人耳来说感觉比较粗糙，频率到梅尔频率的关系曲线如图 3-5 所示。

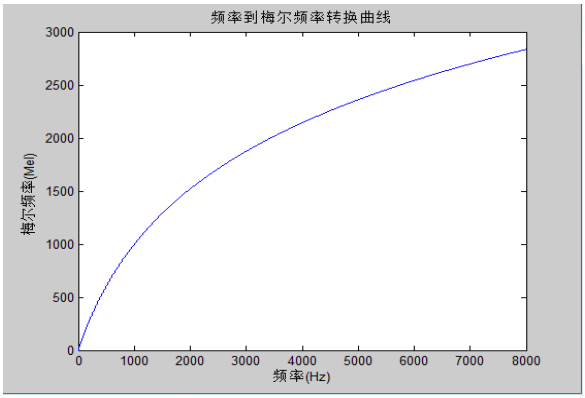



图 3-5 梅尔频率与频率的关系曲线

由于人耳的听觉掩蔽效应，当两个频率比较接近的语音同时发声，会对人耳造成干扰。要想区分两种频率的语音，他们之间必须相隔一定的带宽。这个带宽被称为临界带宽（Critical Bandwidth）。临界带宽的计算公式为：

$$BW_c = 25 + 75[1 + 1.4(f_c / 1000)]^{0.69} \quad (3\text{-}18)$$

根据上述结论可以使用一组临界频带滤波器（Critical Band Filter Bank）对人类的听觉特性进行模仿。临界频带滤波器组的频率分布在梅尔频率刻度上为线性刻度，且滤波器的带宽在临界带宽之内。根据梅尔频率与频率之间的关系在实际应用中的临界带宽与频率的关系可做表 3-1 的近似，通过梅尔频率与频率的关系公式也可以得出类似的结论。

## 3.4 梅尔倒谱频率参数的提取

在各种语音信号处理应用中首先都是对语音信号序列（PCM 码流）进行预处理，它的主要目的是去除静寂音，在判断信号是否为静寂音的时候一般都是在时域进行处理（如可以设定短时能量和短时过零率的门限来判断）。计算语音数据的能量：

$$E = \sum_{n=1}^{N} s^2(n) \quad (3\text{-}19)$$

如果连续语音帧的能量大于预设静音阈值，则保留该段连续语音帧为训练帧，否则舍弃该段语音。

在语音识别和说话人识别中，梅尔（mel）倒频谱系数是运用最多的特征参数。我们可以使用梅尔刻度的滤波器组来处理语音信号。MFCC 的计算公式如 3-20：

$$x_d = \sum_{k=1}^{K} \log(\hat{S}_k \cos\left\{ d\left(k - \frac{1}{2}\right)\frac{\pi}{K} \right\}, d = 1, 2 \ldots D \quad (3\text{-}20)$$

其中 D 是倒谱长度。图 3-6 表示滤波器组的频率分布，根据采样定理，滤波器组最少要覆盖信号频率。每个滤波器都是三角带通滤波器，带宽之间相隔一个固定的梅尔频率刻度。

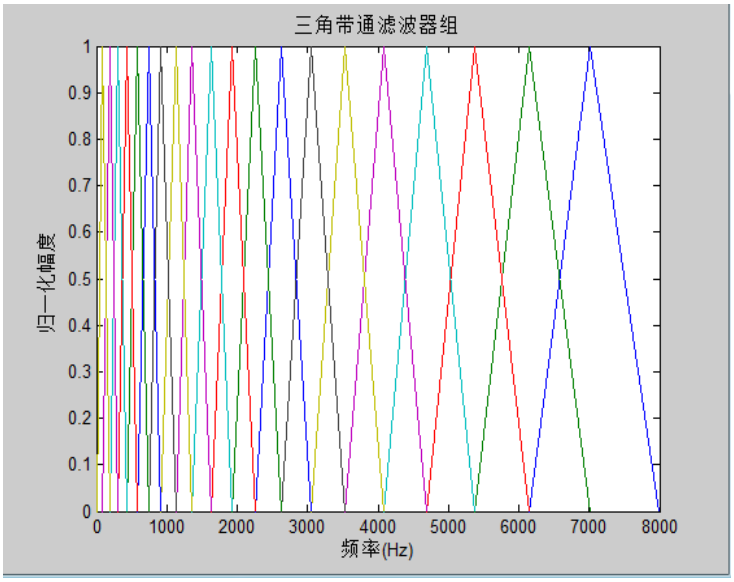



图 3-6 三角滤波器组的频率分布

经过预处理阶段后就可以对语音信号进行梅尔倒谱系数的提取，提取的流程图如图 3-7 所示，大致分为预增强、分帧和加窗、FFT 变换、滤波器组滤波、对数变换、离散余弦变换等几个步骤，下面将详细说明。

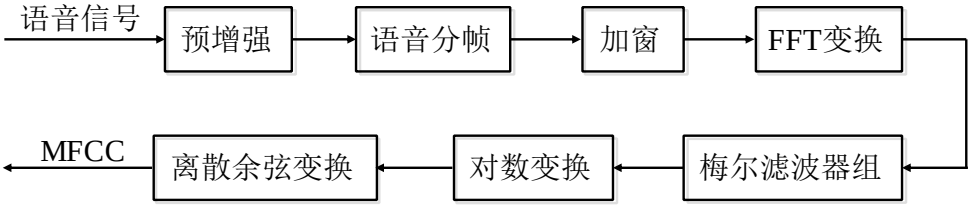



图 3-7 MFCC 提取流程图

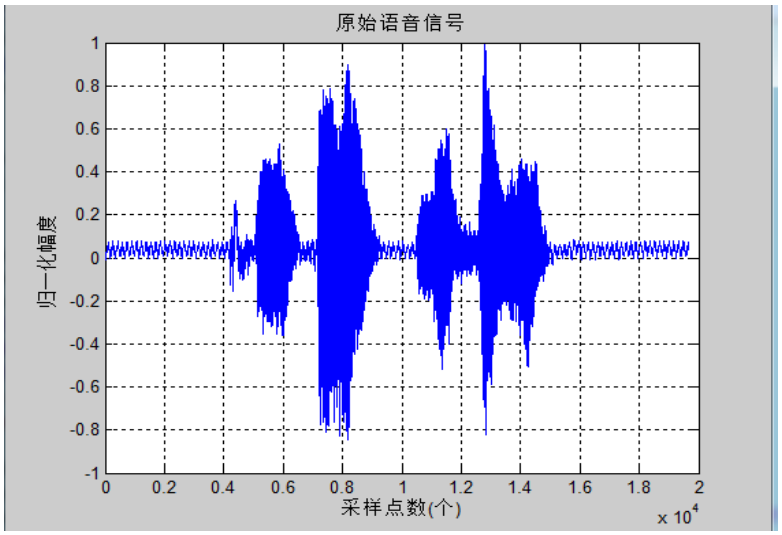



图 3-8 原始语音信号

（1）预增强的功能解释有两种

第一为语音在空气中传送时，高频的能量会随着时间而快速衰减。人耳共振作用可以提高频率段为 2000~5000Hz 的声音强度，刚好可以弥补高频能量的损失，而预增强的作用就是模拟人耳的功能。第二种解释为发声过程中高频能量部分会损失掉，可以用预增强来抵消：

$$H(z)=1-az^{-1} \tag{3-21}$$

其中 H(z)为高通滤波器的 Z 变换，一般会在时域上处理信号。以 S1(n)表示时域的信号，时域预增强公式为：

$$S(n)=S1(n)-a*S1(n)\ \ (0.9<a<1.0) \tag{3-22}$$

本文中的 a 取值 0.95，经过预增强处理后的的语音然需要进行端点检测处理，去除无用的和清音语音段。图 3-9 为经过预加重和端点检测后的语音信号，原始语音信号由图 3-8 给出。

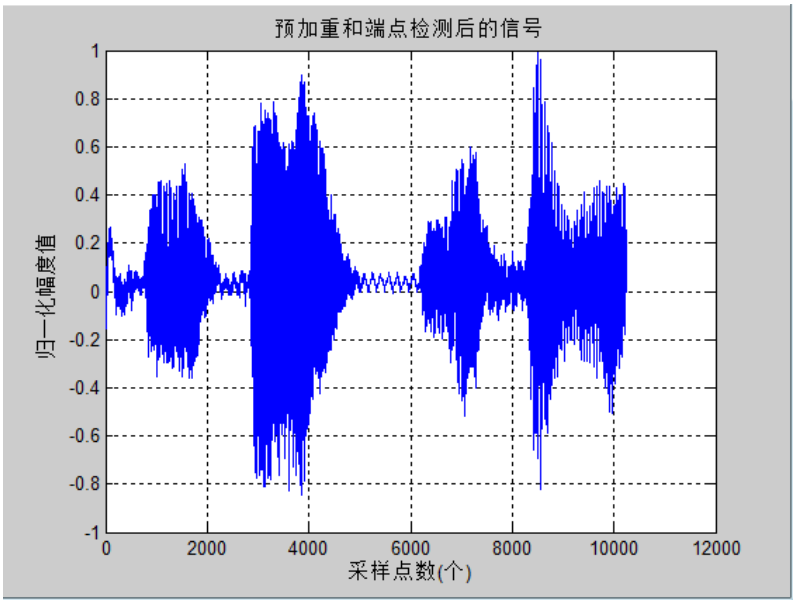



图 3-9 对语音信号进行预加重和端点检测

（2）音框化

即预处理阶段的语音信号分帧。在时域上观察语音信号的波形可以发现，波形的变化十分迅速而且没有一定的规则，但是从频域上观察，则可以发现在短时间段内（20ms~40ms）的情况下频谱是有周期性的。所以在语音信号的预处理中会假设语音信号为短时稳定的，这样我们就可以每隔一小段时间对语音信号取一个音框。

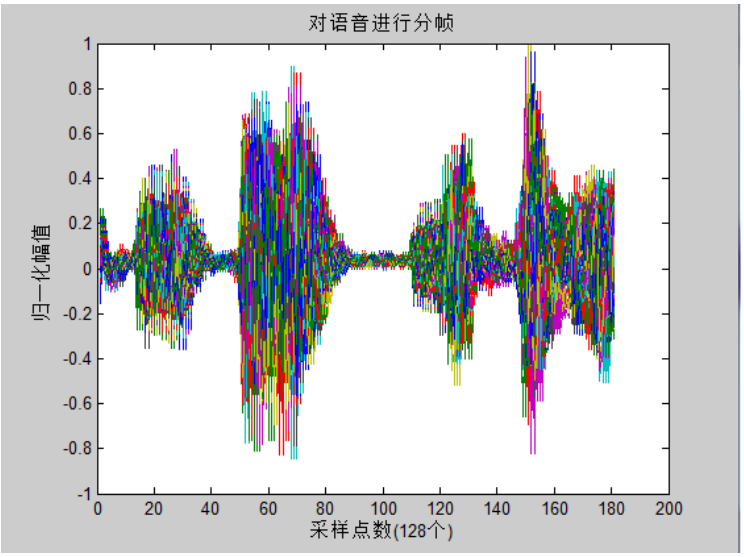



图 3-10 对语音信号进行分帧

为了使各个音框之间有关联，音框与音框之间一般都会重叠一小段时间。例如可以让一个音框长度为 20ms，音框重复时长为 10ms，也因为假设语音信号为短时稳定的，所以就会造成语音信号的连续性遭到破坏，在梅尔倒谱后通常都会进行差分运算。图 3-10 是对语音进行 8000Hz 频率进行取样，窗口时间为 16ms，窗口宽度为 128 来进行分帧。对语音信号的音框化也称为语音信号的分帧处理，分帧处理的下一个步骤就是对分帧后的语音进行加窗处理，其实分帧的语音已经进行过加窗处理，不过所加的窗函数为矩形窗。

（3）加汉明窗

因为每个音框之后都会经过离散傅里叶变换，但是由于每个音框都是固定时间点进行切割，所以就会造成音框的边缘有信号不连续的现象，这就使得音框离散傅里叶变换后会产生高杂信息。

为了降低高杂信息的产生，所以音框在离散傅里叶变换前需要进行加窗函数的处理，以增加音框两端的连续性。将每一个音框（frame）乘上汉明窗，从而增加了音框两端的连续性。设音框化以后的信号（M 帧共 N 点）为 S(n),n =0,1,…..N-1。那么乘上汉明窗后所得到信号为 $S'(n)=S(n)\times W(n)$，其中汉明窗函数为。

$$w(n)=w(n,a)=(1-a)-a\cos\frac{2\pi n}{N-1},0\le n\le N-1,a=0.46 \qquad (3\text{-}23)$$

因为 n 的不同，加窗后即使是混叠的样值也会不同。

图 3-11 为经过前面所述步骤后的语音进行加窗处理后取出的一帧语音段，经过本步骤后的语音滤除了不必要的语音信号，然后就可以进行快速傅里叶变换来

得到语音信号的频谱。

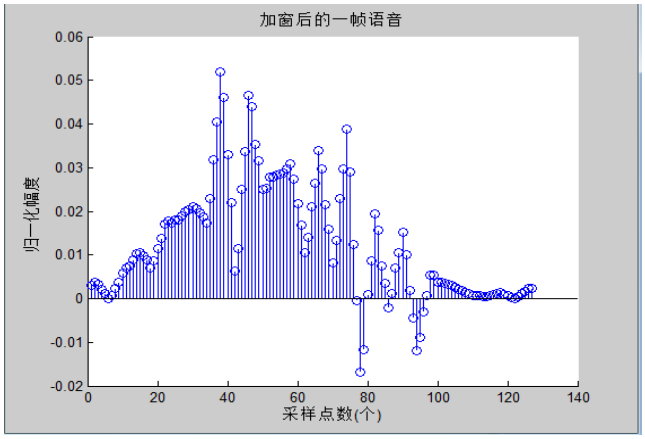



图 3-11 加窗后的一帧语音

（4）快傅里叶变换（FFT）

如前面所述，语音信号在时域上变化迅速且随着时间的改变而不断改变，使得在频域上没办法做有效地观察。但在频域上短时间内语音信号是呈现周期性的。所以一般会将语音信号经过 FFT 变换从时域变换到频域。对 $S'(n)$ 做基 2 的 FFT 变换，不够 2 的倍数补零。快速傅里叶变换公式如下：

$$S'\left(e^{j2\pi k/N}\right)=\sum_{n=0}^{N-1}S_i(n)e^{-j2\pi k/N},n=0,1,\ldots..N-1 \qquad (3\text{-}24)$$

其中 $S_i$ 为第 i 个音框向量，N 为频域上的取样点数。图 3-12 为一帧语音信号经过上述步骤得到 FFT 变换后的频谱图。

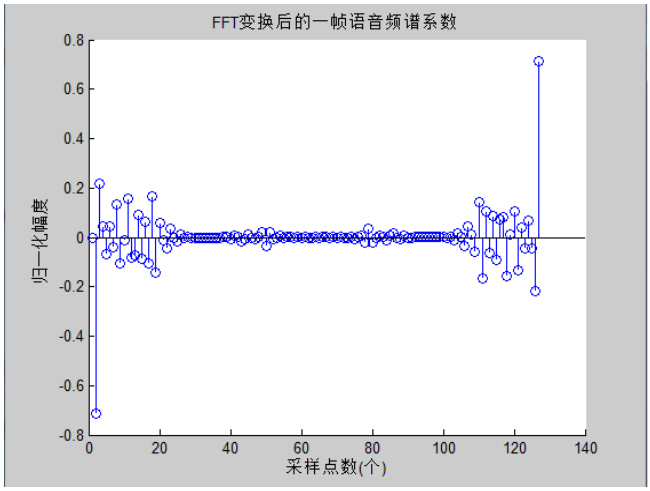



图 3-12 FFT 变换后的一帧语音

（5）三角带通滤波（Triangular Bandpass Filters）

人耳的听觉神经会对某个特定频率范围的声音感觉比较灵敏，但是灵敏度会随着特定频率的偏离而下降。在此特定范围内的频率以对数刻度变化，同时在梅尔频率刻度上成线性形式。所以可以使用 M 个三角带通滤波器组成的滤波器组来模拟人耳听觉特征。需要注意的是所有滤波器要包含整个 0 到耐奎斯特频率范围，并且在梅尔频率上均匀分布。三角滤波器的公式如下：

$$H_i(k)=\begin{cases}0, k<f[i-1]\text{或}k>f[i+1]\\ \dfrac{2(k-f[i-1])}{(f[i+1]-f[i-1])(f[i]-f[i-1])}, f[i-1]\le k\le f[i]\\ \dfrac{2(f[i+1]-k)}{(f[i+1]-f[i-1])(f[i+1]-f[i])}, f[i]\le k\le f[i+1]\end{cases} \quad (3\text{-}25)$$

其中 $f[i]$ 是第 i 个三角滤波器的中心点 $H_i(k)$ 为第 i 个三角滤波器的权重，在 mel 频率刻度上是等间隔的，即满足条件：

$$Mel(f[i+1])-Mel(f[i])=Mel(f[i])-Mel(f[i-1]) \quad (3\text{-}26)$$

$f[i]$ 可以进一步表示成：

$$f[i]=\frac{N}{F_s}Mel^{-1}\left(Mel(f_l)+i\frac{Mel(f_h)-Mel(f_l)}{M+1}\right) \quad (3\text{-}27)$$

其中 $F_s$ 为采样频率， $f_l$ 为三角滤波器组中最低的频率， $f_h$ 为三角滤波器组中最高的频率，M 为三角滤波器组的个数，一般取 M=20。

（6）对数变换

人的听觉会随着频率的升高而越来越不敏感，同时在频域内能量的变化对人耳来说也不敏感。为了模拟人耳的特性，在经过梅尔三角滤波器以后一般还需要对滤波器的输出做对数变换。

（7）离散余弦变换（DCT）

梅尔倒谱系数的求取过程的最后一步就是上一步对数转换的结果再通过离散余弦变换到时域，从而得到梅尔倒谱系数：

$$C_l(n)=\sum_{k=1}^{K}\log|S_i(e^{\frac{j2\pi n}{N}})|\cos(n(k-0.5)\frac{\pi}{k}\quad ,n=0,1,2,...,L \quad (3\text{-}28)$$

其中 $S_i(g)$ 是第 i 个音框向量在频域的成分，N 是频域上的取样点数，n 是第 n 个梅尔倒谱特征。离散余弦变换有两个目的：第一降低维度间的关系，有助于减少数据量。第二，降低了维度，从而加快了识别的效率。

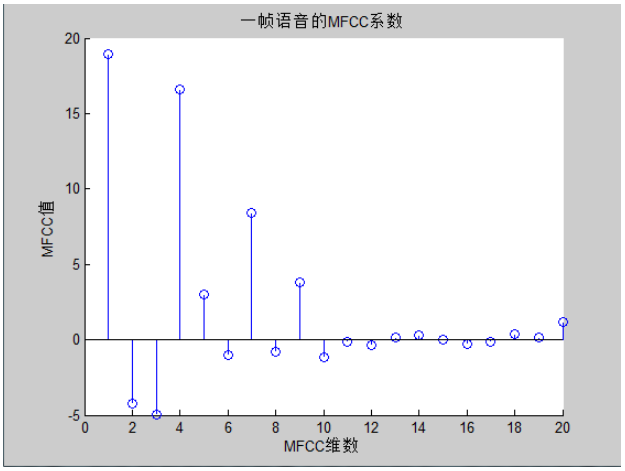



图 3-13 一帧语音的 MFCC 系数

（8）能量及差量

不同的音素，如声母和韵母在能量上差别比较大，由此可知能量也可以作为衡量声音的重要特征，一般会把能量与梅尔倒谱特征结合，公式 3-29 为能量计算公式。其中 $b_i^2$ 代表语音信号通过第 i 个滤波器后的能量，e 代表语音信号通过所有滤波器后的对数能量的组合。

$$e=\sum_{i=1}^{N}\log b_i^2 \tag{3-29}$$

一开始假设语音信号为短时稳定的，所以每隔一段时间取一个音框，这种假设造成了语音信号的破坏，使得音框与音框之间是相互独立的。其实在实际的语音中音框间存在着连续性的关系，为了修正短时稳定的假设带来的负面影响，所以一般在最后进行 MFCC 提取后会加上 MFCC 的一阶差分 $\Delta C_l[n]$ 和二阶差分 $\Delta^2 C_l[n]$ 的差量所得的向量，然后和 MFCC 参数合并。两者的计算公式如下：

$$\Delta C_l[n]=\frac{\sum_{p=1}^{P}p(C_{l+p}[n]-C_{l-p}[n])}{2\sum_{p=1}^{P}p^2} \tag{3-30}$$

$$\Delta C_l^2[n]=\frac{\sum_{p=1}^{P}p(\Delta C_{l+p}[n]-\Delta C_{l-p}[n])}{2\sum_{p=1}^{P}p^2} \tag{3-31}$$

其中 n 为维数（值为 MFCC 维数与一阶二阶差分维数的和，如 MFCC 取 16 维则 n=48）。

## 3.5 梅尔倒谱频率的改进

### 3.5.1 凯泽窗（kaiser windowing）

前面已经提过，传统的 MFCC 的提取过程中所加的窗口函数都是汉明窗，这里我们使用凯泽窗来对语音信号进行加窗处理，其目的是使均方误差降到最小[40]。凯泽窗有一个调整参数 $\alpha$ 控制着波形边缘趋于零的快慢。凯泽窗由下式给出：

$$w(n)=\frac{I_0\left(\alpha\sqrt{1-\left(1-\frac{2n}{N-1}\right)^2}\right)}{I_0(\alpha)},0\le n\le N-1 \qquad (3\text{-}32)$$

其中 $I_0(\alpha)$ 是修正过的 0 阶贝赛尔曲线。$\alpha$ 越大主瓣越宽，其取值依赖于 N[41]。对于相同的 N，凯泽窗可以提供不同的过渡带宽。

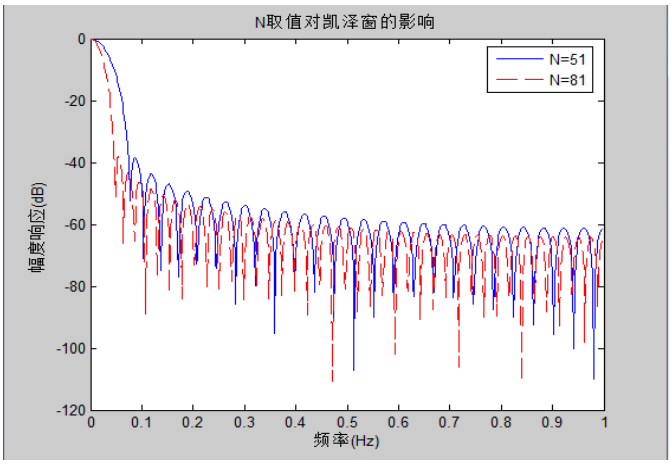



图 3-14 $\alpha$ 对的凯泽窗的影响

### 3.5.2 取 FFT 的绝对值

传统的 MFCC 提取过程中都是将 FFT 运算结果的平方进行梅尔滤波器组滤波，本文将 FFT 运算的结果取绝对值，这样可以大大节省计算时间，提高了系统的效率。

### 3.5.3 加权的 MFCC

在用 MFCC 作为特征参数的建模过程中，与时间相关的动态信息将会丢失。解决这种局限性的最常用策略是添加 MFCC 系数的 Δ（一阶时间导数或速度函数）

和$\Delta^2$（二阶时间倒数或加速度函数）到静态 MFCC 数据或绝对值中来解决。但是这样一来将会增大了运算复杂度。很多 MFCC 的修改方法被提出， MFCC 的 3~18 维包含了说话人信息最丰富[42]，本文提取特征参数 MFCC 的维数为 20 维，分别去掉对贡献率不大的前两维和后两维，得到 16 维特征分量（MFCC），根据各维分量对识别结果的贡献率，然后通过半生正弦函数加权可以得到二次处理后的 MFCC（共 16 维），半升正弦函数如图 3-11 所示。半升正弦函数由公式 3-33 给出，其中 p 为维数。

$$y = 0.5 + 0.5\sin(\frac{\pi i}{p}) \ ,1 \le i \le p \tag{3-33}$$

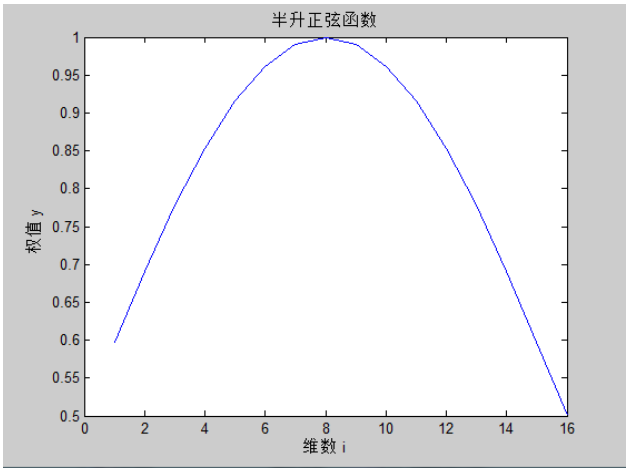



图 3-15 半升正弦函数

本文在此基础上使用了一种简单的加权 MFCC 参数方法在加入动态 MFCC 一阶差分和二阶差分参数的同时对他们进行加权：

$$Wc(n) = c(n) + p\Delta c(n) + q\Delta^2 c(n) \tag{3-34}$$

其中 $p$ 和 $q$ 为加权系数，因为一阶差分和二阶差分参数相比较于 MFCC 参数所起的作用会逐渐的减小，所以令 $q < p < 1$ 则最后的 WMFCC 参数总共有 16 维（未加权以前共有 48 维），从而在保证了识别效果的基础上，减少了数据复杂度。

# 第四章 识别模型的方案与设计

在前面的章节中介绍了语音信号的特征向量的提取方法，说话人识别的另外一个部分就是对提取的特征向量进行建模，建模方法有很多种。本章中我们介绍用的最为广泛的高斯混合模型。高斯分布（Gaussian distribution）又称为正态分布（Normal distribution），是一个在数学、物理及工程等领域都非常重要的连续概率分布函数，它描述了一种围绕某个单值聚集分布的随机变量。在实际生活中，许多物理现象以及各种心理学测试分数都近似地服从高斯分布。在统计学以及许多统计测试中高斯分布也是应用最广泛的一类分布。高斯混合模型（Gaussian mixture model，GMM）是单一高斯密度函数的扩展，由于高斯混合模型可以逼近任意形状的概率密度分布 [43]，所以高斯混合模型被广泛的运用到各种领域，如语音识别，图像识别等，并取得了理想的效果。

## 4.1 单一高斯概率密度函数

假设有一组 D 维概率分布点 $x_i$ ,$i=1,2,....n$，若所有点的分布呈现如图 4-1 的椭球状时，为了描述这些点的产生概率，可以用高斯密度函数来表达其概率密度函数：

$$N(x;\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}\,|\,\Sigma_k\,|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^{'}\Sigma^{-1}\left(x-\mu\right)\} \qquad (4\text{-}1)$$

其中 $\mu$ 代表密度函数的中心点也就是均值，Σ 为密度函数的协方差矩阵，为非奇异的。这些参数决定了此密度函数的特性，如函数形状的中心点，宽窄及走向等，需要注意的是 x 为 D 维的向量[44]。

若要求资料点最佳描述估计参数，可以用最佳可能性估测法的概念来求得。在上述高斯概率密度函数的假设下，当 $x=x_i$ 时，其概率密度为 $N(x_i;\mu,\Sigma)$，若我们假设 $x_i$ ,$i=1,2,....,n$ 之间的各个事件为独立分布事件，则发生 $X=\{x_1,x_2,\ldots,x_n\}$ 的概率密度为：

$$p\left(X;\mu,\Sigma\right)=\prod_{i=1}^{n}N(x_i;\mu,\Sigma) \qquad (4\text{-}2)$$

由于 X 是已经发生的事件，所以我们希望找到 $\mu,\Sigma$ 的值，使得 $p(X;\mu,\Sigma)$ 的值最大，该种参数估计 $\mu,\Sigma$ 值的方法称为最大似然估计（MLE）。要求 $p(X;\mu,\Sigma)$ 的最大值，我们通常将上式两边取对数将乘法变成加法运算来求 $J(\mu,\Sigma)$ 的最大值：

$$J(\mu,\Sigma)=\ln p(X;\mu,\Sigma) \tag{4-3}$$

要求得最佳的 $\mu$ 的估计值，直接求 $J(\mu,\Sigma)$ 对 $\mu$ 的微分即可：

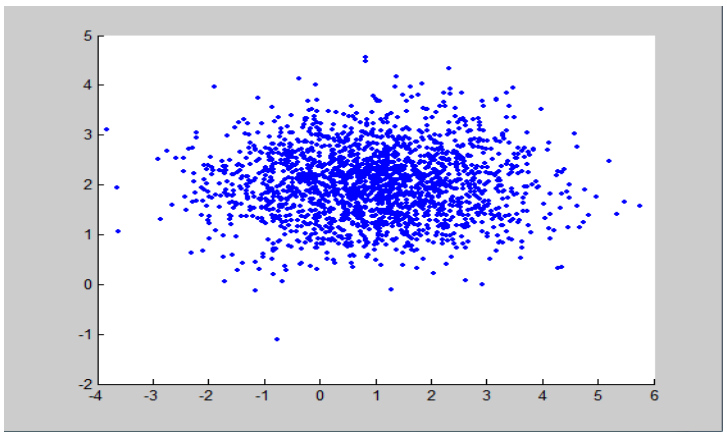



图 4-1 1000 个点的高斯分布

$$\nabla_{\mu}J(\mu,\Sigma)=-\frac{1}{2}\sum_{i=1}^{n}2\Sigma^{-1}(x_i-\mu)=-\Sigma^{-1}(\sum_{i=1}^{n}x_i-n\mu) \tag{4-4}$$

令上式等于零，则可以得到 $\mu$ 的最佳估计值：

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{4-5}$$

同理可以求得 $\Sigma$ 的最佳估计值：

$$\hat{\Sigma}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})' \tag{4-6}$$

## 4. 2 高斯混合密度函数

当数据分布如图 4-2 所示，不再是椭球状的时候，因为单一的高斯分布只有一个极大值，所以就不能用单一的高斯概率分布来进行描述了。通过观察图 4-2 我们发现数据的分布主要集中在三个部分，而每一个部分都可以用一个单一的高斯概率分布来进行描述，这就引出了高斯混合概率密度函数：即用一组单一的高斯概率分布的加权后的线性叠加来进行描述。随着单一高斯混合密度函数的增加，几

乎所有的连续分布都可以用这种方法来进行逼近。

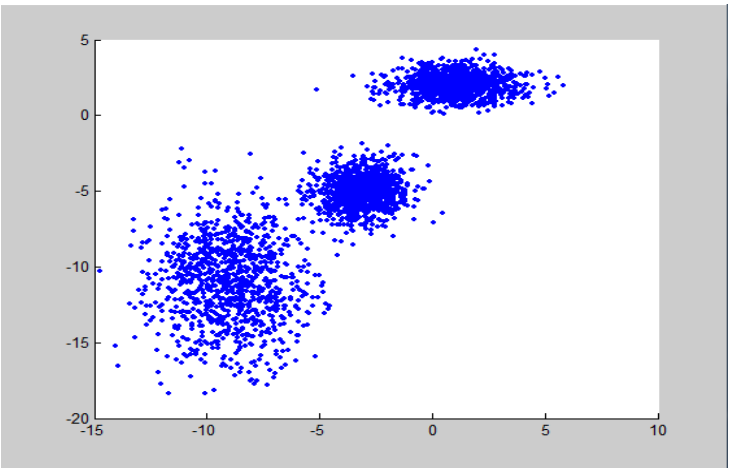



图 4-2 3000 个点的高斯分布

设有 M 个单一高斯概率分布函数，一个高斯混合概率分布可以定义为：

$$p(x)=\sum_{M=1}^{M}p(m)p(x\mid m)=\sum_{m=1}^{M}\pi_m N(x;\mu_m,\Sigma_m) \tag{4-7}$$

其中为$\pi_m$为单个高斯概率密度函数的权值，并且所有权值$(\pi_1,\pi_2,\ldots,\pi_M)$满足关系式：

$$\sum_{m=1}^{M}\pi_m=1\ ,0\le\pi_m\le 1 \tag{4-8}$$

此概率密度函数的参数为$(\pi_1,\pi_2,\ldots,\pi_M,\mu_1,\mu_2,\ldots,\mu_M,\Sigma_1,\Sigma_2,\ldots,\Sigma_M)$，高斯混合模型的组成如图 4-3 所示。

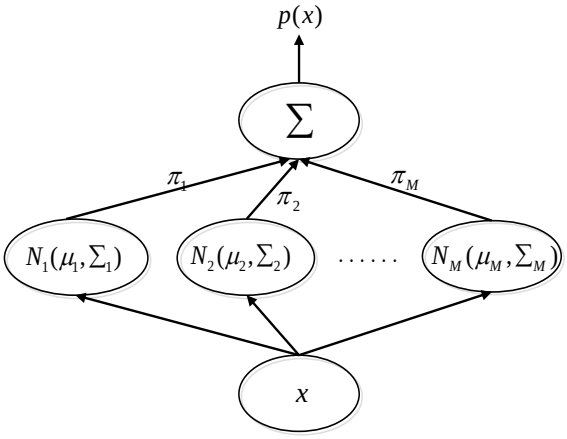



图 4-3 高斯混合模型

从一个高斯混合模型中选取一个点分成两个步骤：首先从 M 个单一高斯分布

中选取一个分布函数，该函数被选中的概率为其加权系数$\pi_m$，第二步再从被选出要找的点；这就转化成为了单一的高斯分布的计算的问题了。为了计算上的方便，通常设各个高斯密度函数的协方差矩阵可以表示为对角阵乘以一个常数的形式:

$$\Sigma_j = \sigma_j^2 I = \sigma_j^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 \\ \mathrm{M} & 0 & \mathrm{O} & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{4-9}$$

此时单一的高斯概率密度函数可以表示如下：

$$N(x;\mu,\sigma^2) = (2\pi)^{-D/2}\sigma^{-D}\exp\left[\frac{(x-\mu)'(x-\mu)}{2\sigma^2}\right] \tag{4-10}$$

在说话人识别中我们可以将语音信号看作一个随机过程，从而对发音的过程建模。我们可以将提取的每帧语音的特征向量看作是一个正态分布，各个帧之间总体上也成正态分布，给每个帧一个权值，从而各个正态分布的组合表征了该说话人的特征。说话人识别首先根据每个说话人的语音进行训练得到一组参数（权重，均值，协方差矩阵）用来表征该说话人的身份，这就需要对高斯混合模型的各个参数进行估计，也称为模型训练。

## 4.3 说话人识别模型训练

在建立说话人的模型中，对于每一个说话人来说需要找出其高斯混合模型参数（均值向量，协方差矩阵，以及混合权值）的初始值，因此先用向量量化（VQ）的方法对训练的语音来求各个初始值，向量量化使用 LBG 算法将所有的语音样本分成若干个类别，并以向量量化所得到的代表点来作为初始值，然后利用期望值最大化算法重新估计模型的各个参数。训练流程如图 4-4 所示。

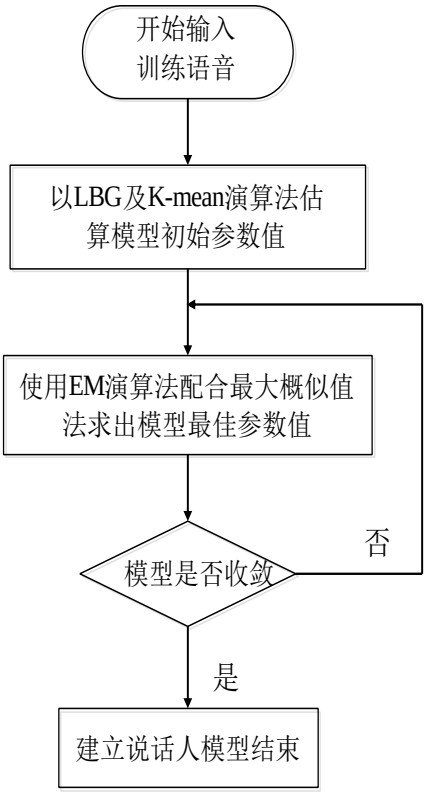



图 4-4 说话人模型训练流程图

## 4.4 模型的参数估计

在进行说话人训练时，我们主要的目标是希望找出最能代表说话人语音特征向量分布的参数集合 $\lambda=[\pi,\mu,\Sigma]$，在实际中使用最多的方法就是最大似然估计法（Maximum likelihood estimation）。最大似然估计法主要是从说话人的语音训练数据中，找出一组高斯混合模型的参数集合 $\lambda$，来使得高斯混合模型的似然值为最大[45]。假设某一训练语句经过特征参数提取后，得到 T 个特征向量，其集合为 $X=\{x_1,...,x_T\}$，则高斯混合模型的概率可写成 4-11 式。

$$p(X|\lambda)=\prod_{t=1}^{T}p(x_t|\lambda) \tag{4-11}$$

上式称为似然函数，由于单个点的概率比较小，他们的乘积可能会造成浮点数溢出，所以通过取其对数形式转化成相加的形式，得到 log 似然函数：

$$\log p(X|\lambda)=\log\prod_{t=1}^{T}p(x_t|\lambda)=\sum_{t=1}^{T}\log\left(\sum_{m=1}^{M}\pi_m N(x;\mu_m,\Sigma_m)\right) \tag{4-12}$$

由于在公式（4-12）中的参数集合$\lambda$是非线性函数（对数里面有加与和），因此我们无法直接利用该式对参数$\lambda$微分令等于零的方式来求解，但是我们仍可利用期望值最大化 （Expectation-maximization, EM） 算法反复地估算最大可能性的高斯混合模型参数，直到收敛为止。

## 4.5 EM 算法

为了求得 GMM 模型的参数，一般采用期望值最大算法来估计模型参数，EM 算法的主要目的就是找到一个$\lambda$使得$p(X|\lambda)$最大，具体作法是先找一个初始模型的参数$\lambda$来估算新的模型参数$\lambda'$，使得$p(X|\lambda') \geq p(X|\lambda)$。然后新的模型参数$\lambda'$变成初始模型参数$\lambda$，反复的重复此步骤，直到$p(X|\lambda)$收敛为止[46]。可以采用在高斯混合模型中使用的随机选点的两步法来进行解决。

（1）E-step：估计每个单一高斯概率密度函数生成的概率（并不是指每个单一高斯概率密度函数被选中的概率）。对于数据$X=\{x_1,...,x_T\}$中的$x_i$来说由第 j 个单一高斯概率密度函数生成的概率即$\pi_j$的后验概率为：

$$\beta_j = E(\pi_j | x_i;\lambda) = \frac{\pi_j N(x_i|\mu_j,\Sigma_j)}{\sum_{i=1}^{M}\pi_j N(x_i|\mu_j,\Sigma_j)} \quad ,1\leq i\leq n,1\leq j\leq M \qquad (4\text{-}13)$$

在上式中的均值和协方差矩阵也是需要估计的，所以可以采用迭代法：假定它们为已知的，通常取上一次的迭代值，这就需要一个初始值（后面做介绍）。在实现时通过$N_j(x;\mu_j,\Sigma)$计算$x_i$在高斯模型中的概率可以得到一个 n 行 1 列的向量，最后得到一个$n\times M$的矩阵，则矩阵的列向量为所有点在概率模型下的概率。

（2）最大化（M-step）：估计每个单一高斯概率分布的参数，假设在 E-step 中计算得到的后验概率$\beta_j$就是数据$x_i$由第 j 个单一高斯概率分布生成的概率，因为每一个单一高斯概率分布都是一个标准的高斯分布，这样就可以得到最大似然的参数值：

更新权值：
$$\pi_j' = \frac{\sum_{i=1}^{T}\beta_{ij}}{N} \qquad (4\text{-}14)$$

更新均值：
$$\mu_j^{'} = \frac{\sum_{i=1}^{T} \beta_{ij} x_i}{\sum_{i=1}^{T} \beta_{ij}} \tag{4-15}$$

更新协方差矩阵：
$$\Sigma_j^{'} = \frac{\sum_{i=1}^{N} \beta_{ij} (x_i - \mu_j^{'})(x_i - \mu_j^{'})^{'}}{\sum_{i=1}^{N} \beta_{ij}} \tag{4-16}$$

不断的进行迭代，重复上面的步骤，直到新旧参数相差一个很小的正数$\left|\lambda - \lambda^{'}\right| < \varepsilon$。

## 4.6 EM 算法的初始化

使用 EM 算法首先要获得高斯混合模型的阶数 M 和初始化参数 $\lambda_0$。常用的初始化方法可以分为以下几种。

（1）随机选择法：均值 $\mu_i, i = 1,2,3...,M$ 的初始值为语音数据中随机选取 M 个向量，$\Sigma_i, i = 1,2,\ldots M$ 初始化为单位阵。

（2）平均分段法，将语音均分为 M 段，求各个分段的平均值，然后根据均值来求出方差，权重初始化为为$1/M$。

（3）聚类选择法:聚类法用的最多的就是 LBG 以及 K-means 算法，这种方法与高斯混合模型的组成原理相吻合，根据样本先验概率分布的理论，将特征矢量分为 M 个聚类，进行初始化。

### 4.6.1 K-均值算法

如上面介绍，K-means 算法是聚类的一种，通过最小化各个点到中心点的距离的平方和来完成，具体初始化步骤为[47]：

（1）初始聚类中心由 M 个矢量组成，这些矢量可以随意选取，为了简单我们选取前 M 个矢量$(z_1(1), z_2(1), ..., z_M(1))$；

（2）数据共有 M 个聚类，对于一个样本点$x_k$，由最小距离的准则，计算出距离最小的那个聚类，计算公式如下：

$$|x_k - z_i(m)\leq |x_k - z_j (n \quad \forall i \neq j \text{ 并且 } i, j = 1,2,...,M \tag{4-17}$$

从而将$x_k$归于第$i$类$S_k$；

（3） 更新聚类中心，计算公式有下式给出，其中$N_i$为聚类$S_i$中的样本个数

$$z_i(m+1)=\frac{1}{N_i}\sum_{x_k\in z_i(m)}x_k \qquad i=1,2,\ldots M. \tag{4-18}$$

（4） 若$|z_i(m+1)-z_i(m)|\geq\delta$则转到（2）继续。

（5） 初始化 GMM 模型的各个参数，总样本个数为 T

$$\alpha_i=\frac{N_i}{T} \tag{4-19}$$

$$\mu_i=\frac{1}{N_i}\sum_{x_k\in z_i}x_k \tag{4-20}$$

$$\sigma_{ik}^2=\frac{1}{N}\sum_{x_k\in c_i}(x_{ik}-\mu_{ik})^2 \quad ,k=0,1,\ldots,D-1 \tag{4-21}$$

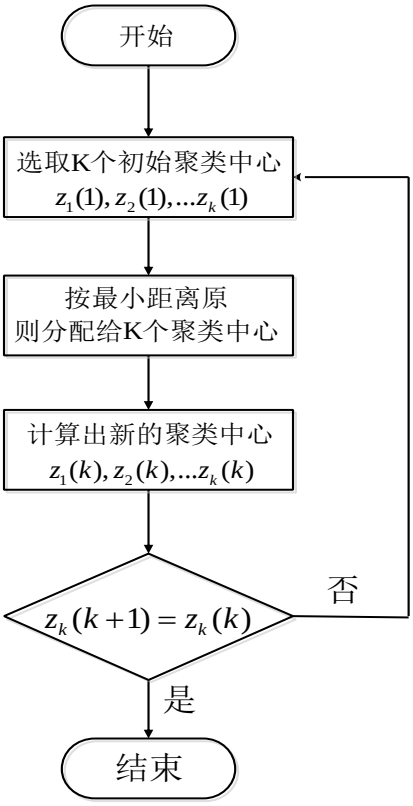



图 4-5 K-均值算法流程图

图 4-5 给出了 K-均值算法的流程图，由于算法开始的时候用到的中心值为随机选取的，如果开始选取的值不理想，就可能造成聚类结果的局部最优，从而得不到想要的结果。但是由于 K-均值算法比较容易实现，在实际应用中依然比较广泛，通常将 K-均值算法和其他算法进行结合来实现最优结果，例如和 LBG 算法的结合来进行聚类，下面将进行介绍。

### 4.6.2 LBG 算法

VQ 被广泛的应用到了数据压缩的领域，其根本原理就是将看似没有关系的一组随机向量集中成几个具有代表性的点，来代表样本点在向量空间中的位置。利用向量量化可以将大量的语音样本分类，但是不能讲样本在向量空间的分布形状及大小描述出来。因此常常和高斯混合模型相结合。LBG[48]算法来求高斯混合模型参数的初始值，其步骤和可以分成六步。

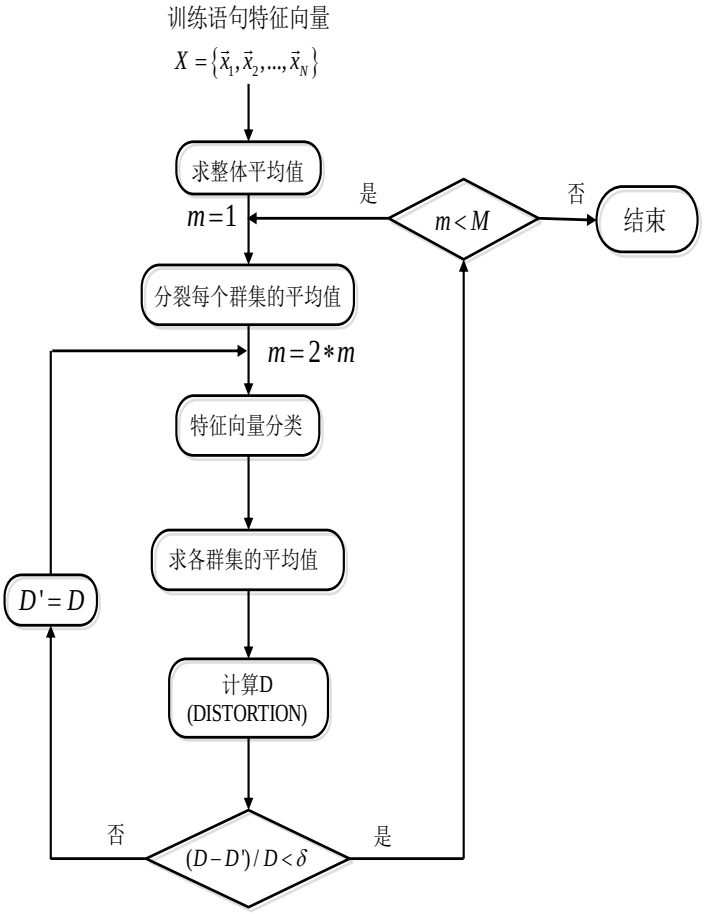



图 4-6 LBG 算法流程图

（1）求整体特征向量的平均值，当作起始值：

$$\mu=\frac{1}{T}\sum_{t=1}^{T}\mathrm{r}\,x_t \tag{4-22}$$

（2）将每一个平均值分裂成两个，$\varepsilon$ 为分裂系数，可以取为 0.01。

$$\begin{aligned}\mathrm{r}\,\mu_m^{+}&=\mathrm{r}\,\mu_m(1-\varepsilon)\\ \mathrm{r}\,\mu_m^{-}&=\mathrm{r}\,\mu_m(1+\varepsilon)\end{aligned} \tag{4-23}$$

（3）利用 k-mean 算法将所有的特征向量按照步骤 2 分裂出的平均值 $\overset{1}{\mu}_m$ 进行重新分裂。

（4）更新平均值 $\overset{1}{\mu}_m$，按照步骤 3 的结果计算出每个群集新的平均值。

（5）重复步骤 3,4 直到每一个平均值与特征向量的总体距离和获得最小的失真也就是当更新率小于 $\delta$ 时停止，本文将 $\delta$ 设为 0.01。

（6）重复 2，3，4，5 步骤直到分裂值所设定的数目。

## 4.7 说话人识别的判决法则

在第一章中已经介绍过说话人识别分为说话人确认和说话人辨认，说话人识别的最后一个步骤就是判决过程，下面介绍一下说话人确认和说话人辨认的判断准则。

### 4.7.1 说话人辨认

在声学特性上每个人的声音特征不同，所以在与文本无关的说话人识别中，每个人的训练模型都包含自己独有的语音特征，而说话人识别就是讲测试者的语音模型和已经建模的语音库中的人进行对比，并根据似然函数 $p(X \mid \lambda_k)$ 从中挑选出最相似的说话人模型，具体过程如下：

对于 S 个说话人，我们以高斯混合模型 $\lambda_1, \lambda_2, ...., \lambda_S$ 来代表。对于一段测试语音 $X = \{\overset{1}{x}_1, \overset{1}{x}_2, ..., \overset{1}{x}_t\}$，对比已有的高斯混合模型试图找到一个最大后验概率值的模型：

$$\hat{s} = \arg \max_{1 \le k \le S} p(\lambda_k \mid X) \tag{4-24}$$

其中 $\hat{s}$ 表示识别出的说话人，根据最大后验概率与贝叶斯定理上式可以改写为：

$$\hat{s} = \arg \max_{1 \le k \le S} \frac{p(X \mid \lambda_k) p(\lambda_k)}{p(X)} = \arg \max_{1 \le k \le S} \frac{p(X \mid \lambda_k) p(\lambda_k)}{\sum_{M=1}^{S} p(X \mid \lambda_m) p(\lambda_m)} \tag{4-25}$$

假设每个说话人模型出现的概率相同（都为 1/S）则上式可以简化近似为：

$$\hat{s} = \arg \max_{1 \le k \le S} p(X \mid \lambda_k) \tag{4-26}$$

此时最大后验概率就变为了最大似然估计，为了简化运算可以将上式两边进行取对数：

$$\hat{s} = \arg\max_{1\le k\le S} \sum_{t=1}^{T} \log p(x_t \mid \lambda_k) \tag{4-27}$$

### 4.7.2 说话人确认

说话人确认是要对一个宣称的说话人做身份验证[49]，不过系统中要存在这个宣称人，确认的过程就是计算输入的语音与被宣称的说话人模型是否有足够大的相似度。若宣称人的模型参数为 $\lambda_0$，计算输入语音序列 $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_t\}$ 对于此模型相似度 $p(x \mid \lambda_0)$。然后利用除这个宣称人以外的说话人建立一个竞争者模型，同样计算相似度 $p(x \mid \lambda_{race})$，定义对数相似度比值函数：

$$\log p(x \mid \lambda_0) - \log p(x \mid \lambda_{race}) \begin{cases} \ge \theta, \text{接受} \\ \le \theta, \text{拒绝} \end{cases} \tag{4-28}$$

利用对数相似度比值可以将宣称之与其他人的差异拉大。在说话人辨认中常用的一个判决标准是等错误概率曲线，由于本文主要关注说话人辨认，这里不再做详细介绍。

## 4.8 改进的 GMM 模型

在前面的 GMM 模型的参数估计初始化的时候用到了聚类中的 VQ（Vector Quantization）方法，即矢量量化聚类方法，矢量量化在数据压缩和信号处理中被广泛的应用，特别是视频和图像的压缩中都用到了 VQ 方法。VQ 的一种简单的定义是：将向量空间中点用其子向量空间来进行表示。如前面的步骤用一组数据的中心点来代表这一组数据。在说话人识别中，VQ 算法是测量提取的语音特征向量的相似度来对语音向量进行划分，由于 VQ 算法对语音向量数据不需要做任何假设就可以进行建模，所以被广泛应用。VQ 算法在 2007 年以后被广泛用到了语音信号的分析中[48]。

高斯混合模型是一种参数化模型，VQ 是一种非参数模型，从 2009 年开始针对两者的研究越来越多，但是两种方法都存在着一定的不足，为了克服它们的缺点，本文将两者结合提出一种 VQ/GMM 混合模型，虽然 VQ 在识别方面不如 GMM 但是 VQ 模型计算量小。所以我们试图根据 VQ 的优点将他们结合起来，根据说话人模型的相似度来对训练的说话人模型进行分类。

（1）首先我们使用 LBG 算法将提取到的特征向量进行分类（假设分成 N 类），

在分类的时候记录每个子分类的中心均值向量，这些均值向量在识别的时候可以用到。

（2）在训练阶段，根据每个字分类的特征向量的多少来分配一个高斯阶数。这样就可以得到 N 个子 GMM 模型。在这一步骤中需要对每一个训练语音进行分类，分类的标准就是中心均值向量，在第一步中已经得出，整个过程如图 4-7 所示。

（3）测试阶段，将测试语音向量的中心均值向量和每一个子模型的中心向量对比，按照距离最小的原则进行划分。然后将划分后的测试语音向量进行模型对比来找出最大的似然值，然后在进行判定是否为正确的说话人[54]。测试过程如图 4-8 所示。

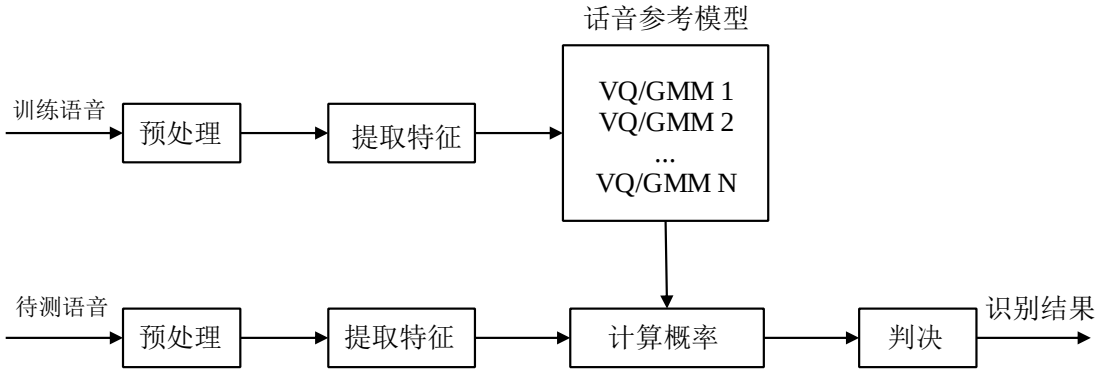



图 4-7 VQ/GMM 说话人识别系统框架图

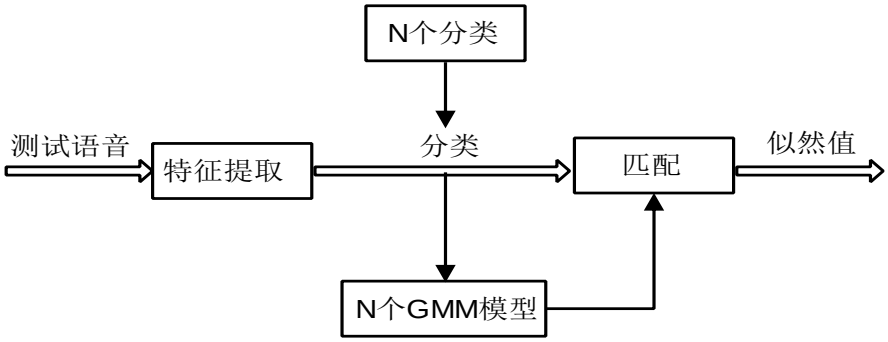



图 4-8 VQ/GMM 测试流程图

当测试人数较多的时候，每个说话人都要进行建模，测试的时候提取的测试语音要和每一个说话人的高斯混合模型进行对比来计算似然值，找出最大的那个来进行判决，能否找到一种方法来减少测试的计算时间，一种很自然的想法就是利用 VQ 先对模型进行分类，分类的方法前面已经介绍。

由于男声与与女声之间存在一定的差异，本文的使用的方法为将高斯混合模型分为男性和女性两个部分。对一个测试模型库来说男生和女生出现的概率大概

相同，所以理论上说，测试时间可以减少一半。具体的流程图如图 4-9 所示。

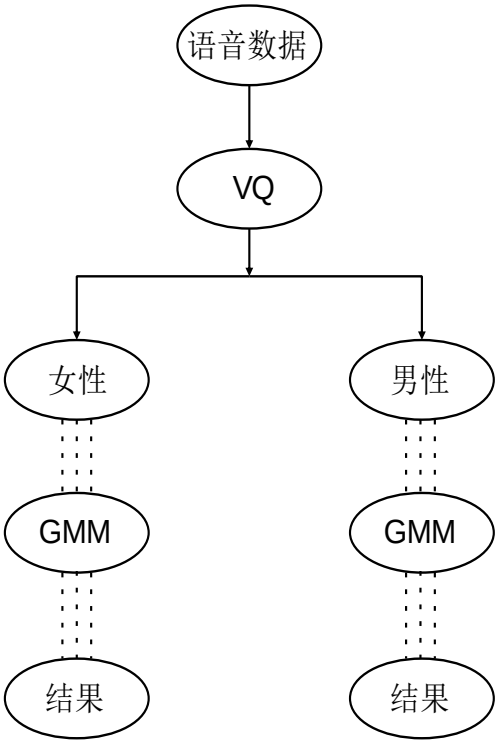



图 4-9 使用 VQ 对 GMM 进行分类

## 4.9 识别率改进方案

在传统的说话人识别系统中，正确的说话人在部分的音框会呈现较低的分数，虽然正确的说话人在其他音框的分数仍高于冒充的说话人，但是最后计算总分时，常因为那一小部分音框而呈现较低的结果，因为总的计算结果是各个结果乘积，所以可能导致总分低于其他说话人，从而造成了识别结果的下降。这些呈现较低分数的音框，可能是有杂音的音框或者是可说话人模型无法匹配有所差异的音框[50]。为了解决这个问题本文首先利用帧投票来对不正常音框进行“去除”，在这个基础上再和传统的 GMM 结合来提高识别率，下面是该方案的详细过程。

如图 4-10 所示，在做说话人识别的时候，每段测试语音最少都有 5 到 6 个呈现较低分数的音框。我们尝试着将这些音框人工移除，那么原本测试错误的测试语音将能够正确识别。为了排除这些可能影响，提出了投票的方法。每个音框都段一票在投票算法中，每个音框都是位单独的一类。

观察说话人识别中的高斯混合模型估计等式：

$$\hat{S} = \arg\max_{1\le k\le M} p(\vec{X} \mid \lambda_k) = \arg\max_{1\le k\le M}[\prod_{t=1}^{T} p(\vec{x}_t \mid \lambda_k)] \quad (4\text{-}29)$$

其中 $\vec{X}$ 为一组语音特征向量$\{x_1, x_2, x_3, ...., x_t\}$。在传统的 GMM 评估中，我们使用 GMM 模型的输出概率几何平均值来找到正确的说话人（即选择均值最大的那个模型所对应的人为正确的说话人）。

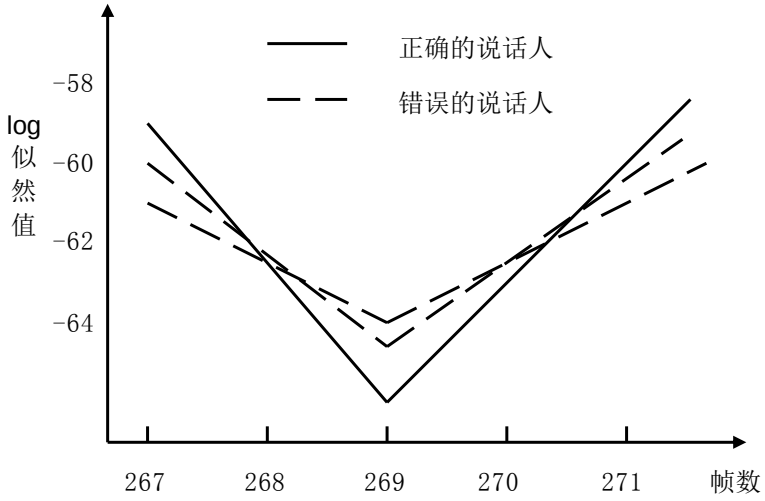



图 4-10 语音中存在分数较低的帧

说话人识别问题可以看做一段语音中的每一帧是一个独立的类。使用 GMM 参数每一个类都独立的判决说话人是谁。在传统的 GMM 模型中，使用各个帧的概率乘积 $p(\vec{x}_t \mid \lambda_k)$ 来进行判断，因为对于一个说话人来说有一些特定的语音帧都每个说话人的判断结果都很小（如噪音和清音），这就影响了最终的判决结果。我们试图找到一种方法来给每帧语音一个权值。在本文中使用帧间的投票而不是他们乘积来进行判决，他们的区别如图 4-8 所示。

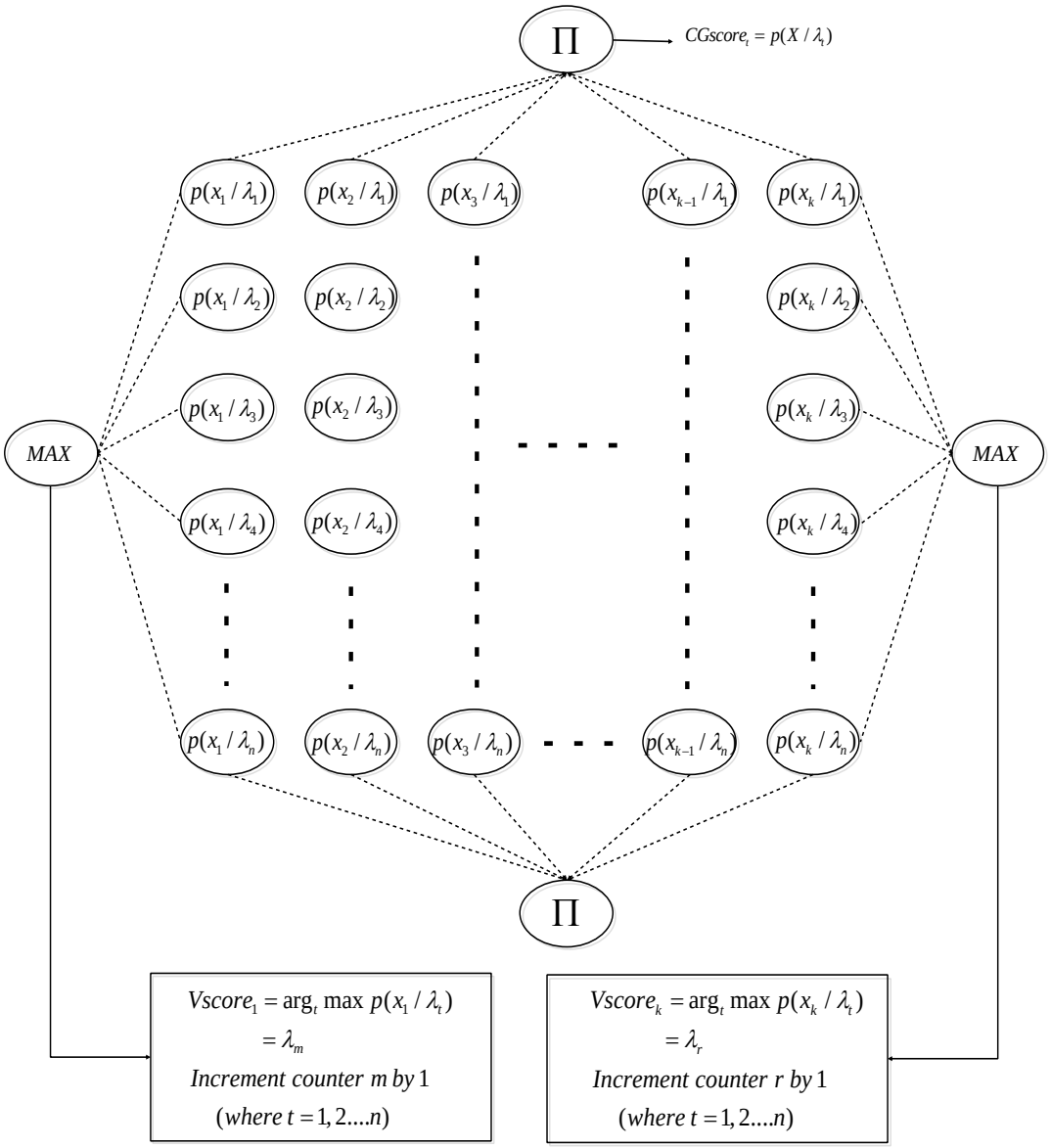
对每一个语音帧都有一个最接近的说话人 $\hat{S}$ 有：

$$\hat{S} = \arg\max_{1\le k\le M} p(\vec{x}_i \mid \lambda_k) \quad (4\text{-}30)$$

对于所有语音帧来说，每一个语音帧都进行一次投票来找到最可能的说话人。判决结果即是得票最多的说话人。这就解决了因为一小部分不好的语音帧而造成的判决失误。算法的伪代码如下：

Initialize a counter for each speaker to 0

For each frame j (loop 1)




$$\text{Recognized speaker } CGscore = \arg_t \max(CGscore_t), t = 1, 2, ...n$$

$$\text{Recognized speaker } Vscore = \arg_s \max(CGscore_s), s = 1, 2, ...k$$

图 4-11 投票法和传统的 GMM 方法的区别

For each speaker i (loop 2)

Evaluate

$$p(\vec{x}_j \mid \lambda_i) = \sum_{k=1}^{M} p_k b_k(\vec{x}_j)$$

End for (loop 2)

Find the speaker v with maximum probability for frame j

$$v = \arg\max_{1\le k\le M} p(\vec{x}_j \mid \lambda_i)$$

v=v+1

end for (loop 1)

the speaker with the largest counter (i.e. largest number of votes) is the correct spearker

帧投票算法的判决等式可以表示为：

$$\hat{S} = \arg\max_{1\le k\le M} \sum_{t=1}^{T} \upsilon(k \mid x_t) \tag{4-31}$$

其中

$$\upsilon(k \mid x_t) = \begin{cases} 1, p(\vec{x}_t \mid \lambda_k) \ge p(\vec{x}_t \mid \lambda_k) \\ 0, p(\vec{x}_t \mid \lambda_k) \ge p(\vec{x}_t \mid \lambda_k) \end{cases} \tag{4-32}$$

在大部分情况下基于帧投票的方法在说话人识别中有着很好的效果，但是在某些情况下如测试的语音有限或者相似的说话人比较多，投票方法的效果往往不尽如人意，本文提出一种方法将两者结合。

本文使用的方法是将竞争者的数目在第一轮中减少，第二轮中再利用帧投票的方法，同时为了不增加计算量，第二轮中不再计算整个语音序列的概率。在第一轮中使用 GMM 模型的方法选出可能性最高的 N 个说话人（特征参数 $\lambda_j$ ）有着最高的 N 个概率 $p(X \mid \lambda_j)$，X 为测试语音序列，作为粗略估计。每个说话人的语音序列的每一帧的概率使用 GMM 方法计算出并存储起来为后续步骤所使用。最高的 N 各说话人然后使用帧投票方法找到得票最高的那个说话人。第二轮中的每一帧的概率在第一轮中已经计算过，所以没有增加计算量。从而很好的解决了人数增加后而造成的不准确问题。

## 4.10 DTW 算法

动态时间规整（Dynamic Time Warping ，DTW）匹配算法是基于动态规划的思想，解决了发音长短不一的匹配问题，它是一种结合了时间规整和距离测度计算的非线性规正技术，是语音识别中出现较早、较为经典的一种算法[53]。

设测试的语音参数有 I 帧特征向量，即测试语音模板的特征矢量序列为 $X = (X_1, X_2, ..., X_I)$，参考语音参数共有 J 帧，则参考模板的特征矢量序列为 $Y = (Y_1, Y_2, ..., Y_J)$ 且 $I \neq J$，则动态时间规整就是要找到一个时间规整函数 $j = w(i)$，将测试矢量的时间轴 i 非线性地映射到参考模板的时问轴 j 上[44]，并使该函数 w 满

足下式：

$$D=\min_{w(i)}\sum_{i=1}^{I}d[X_i,Y_{w(i)}] \tag{4-33}$$

其中 $d[X_i,Y_{w(i)}]$ 是第 $i$ 帧测试矢量 $X_i$ 和第 $j$ 帧模板矢量 $Y_j$ 之间的距离测度，一般这个距离测度采用欧氏距离的平方[10]，如 4-34 式所示。d 则是处于最优时间规整情况下两矢量的距离。

$$d(X_i,Y_j)=\sum_{n=1}^{N}(x_{i,n}-y_{j,n})^2 \tag{4-34}$$

其中 $X_i=(x_{i,1},x_{i,2},x_{i,3},...,x_{i,N})$，$Y_j=(y_{j,1},y_{j,2},y_{j,3},...,y_{j,N})$，$N$ 是特征矢量维数。实际应用中，DTW 一般采用动态规划技术(DP)来实现动态规划是一种最优化算法，其原理如图 4-12 所示。将测试模板的各帧 $i=1,2,...,I$ 作为二维直角坐标系的横轴，参考模板的各帧 $j=1,2,...J$ 作为纵轴。通常规整函数 $w(i)$ 被限制在一个平行四边形内，如图 4-12，它的一条边的斜率为 2，另一条边的斜率为 $1/2$。规整函数的起始点为 $(1,1)$，终止点为 $(I,J)$，即 $w(1)=1$，$w(I)=J$。$w(i)$ 的斜率为 0、l 或 2；否则就为 l 或 2。这是一种简单的局部路径限制。求最佳路径问题可以归结为满足局部路径约束条件，使得沿路径的累积距离最小。这个最小累积距离即为测试语音模板与参考模板语音之间的距离。则与测试模板距离最小的参考模板对应的说话人即为识别结果。本文在测试时使用了 DTW 与 GMM 的结合来进行说话人识别，将两者结合以后的识别率要优于单独使用 GMM 模型。

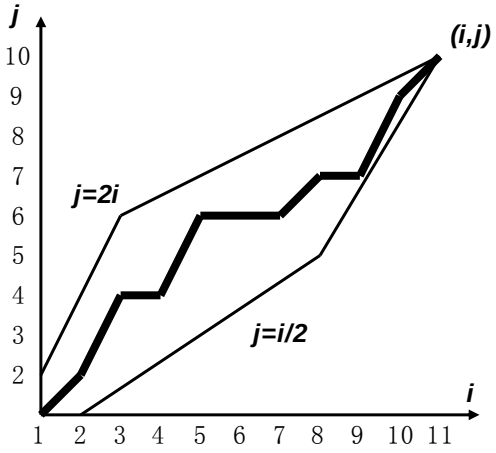



图 4-12 DTW 算法图

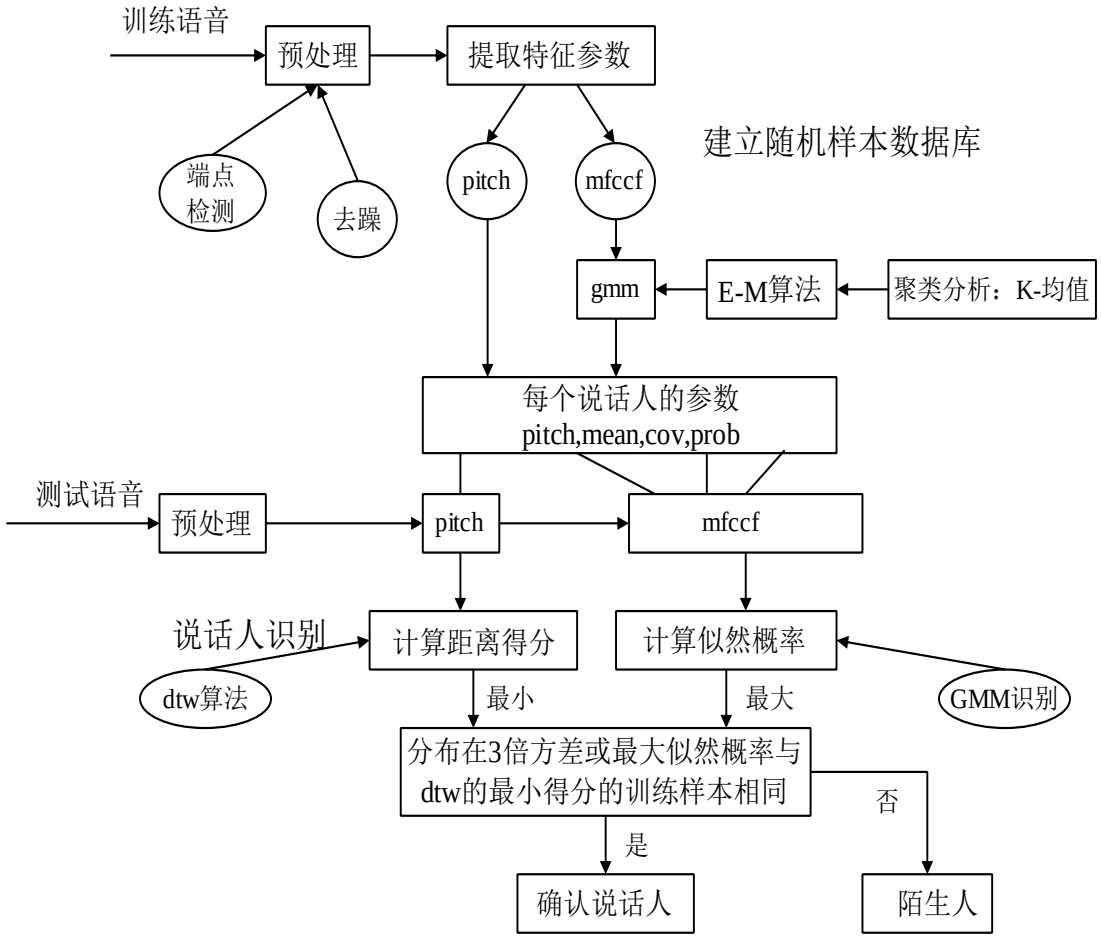



图 4-13 使用 DTW 算法的说话人识别

为了提高系统识别的识别率，将 DTW 和 GMM 相互结合作为说话人识别模型。该过程可以分成两个步骤，首先，用 DTW 算法计算测试样本的 pitch 与各训练样本 pitch 间的最小距离，取得分小的前一部分值（凭数据经验设定）训练样本，然后，用 GMM 计算测试样本特征参数分别在这一部分的训练样本中分布的最大似然概率，对得出的最大似然概率的那个说话人进行判决，最后得出结果。整个流程图如 4-13 所示。该过程通过第一步来减少匹配的数据量，第二步来进行 GMM 识别，从而减少了运算量。

# 第五章 系统实现和实验结果

## 5.1 实验条件

本文的实验基于 Matlab 编程来实现说话人识别，PC 参数为 Windows 7 操作系统，英特尔 2.67GHz 双核处理器，录音采用的笔记本自带的麦克风。

## 5.2 实验语音库

实验所用的语音数据库是由十个男生和十个女生的录音组成，录音环境为普通实验室环境。录音时不限制每个说话人所说的内容，每个说话人都录制了 5 段语音样本，其中语音段时长为 120 秒的一段用于训练样本 4 段用于测试样本，语音段时长 2 到 20s 不等，以 8kHz 频率进行采样，量化位数设置为 16 位，所有语音都保存为通用的 wav 格式。

## 5.3 基于 GMM 说话人识别系统架构

回顾前面章节的介绍，我们知道说话人识别主要分为两个阶段：即训练阶段和测试阶段。其中关键的问题是对语音信号的特征提取以及对语音信号建立模型，图 5-1 到 5-3 给出了说话人识别过程的系统架构分为三个部分：即特征提取，训练和测试。

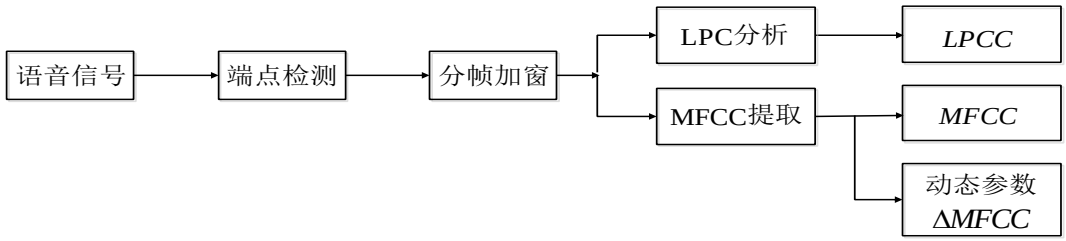



图 5-1 特征提取过程


图 5-2 说话人识别模型训练

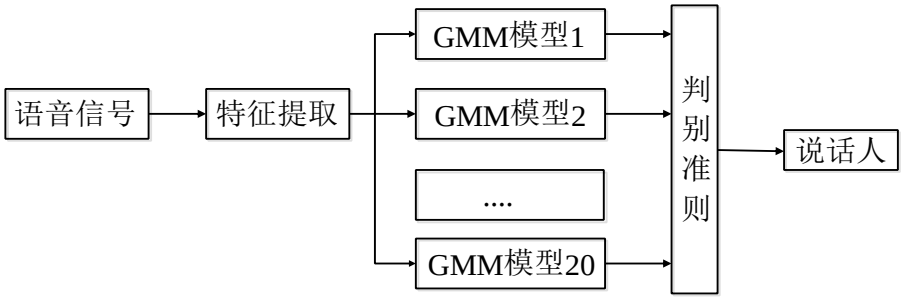



图 5-3 GMM 模型的识别过程

## 5.4 识别率计算

在测试阶段中，对说话人的语音数据提取特征参数以后，可以得到特征向量集合 $X=\{x_1,x_2,...,x_T,x_{T+1},x_{T+2},...\}$，其中每一个特征向量设为 D 维，将特征序列按照下面的方式进行分段，段长度设为 T：

$$\underbrace{x_1,x_2,...,x_T}_{分段1},x_{T+1},x_{T+2},...$$

$$x_1,x_2,...,x_T,\underbrace{x_{T+1},x_{T+2},...}_{分段2}$$

对每一个分段来说都可以看做成一个测试语句。

由前面对于“系统判别准则”的分析，系统识别过程就是将上一步中的各个分段以后，由式（4-42）的判断准则找出在各个模型中计算结果的最大值。然后将这个分段归于该模型。如果最大值模型属于的说话人跟测试语音所属的说话人是同一个，则说明该段测试结果正确，否则说明该分段结果错误[51]。将所有的正确分段数统计，识别率的定义如公式 5-1 所示。

$$正确率(\%)=\frac{正确的分段数}{所有的分段数}\times 100\ (\%) \qquad (5\text{-}1)$$

说话人识别的识别率与段长有关，一般情况下随着 T 的增大识别率会随之升高。

## 5.5 系统实现和程序设计

界面程序的关键代码如下面所示，主要包含了录音模块、将录音数据经过训练后得到语音信号的特征向量并添加到数据库中模块、以及测试等模块。训练以及测试过程的数据将在 matlab 的命令行中进行显示。

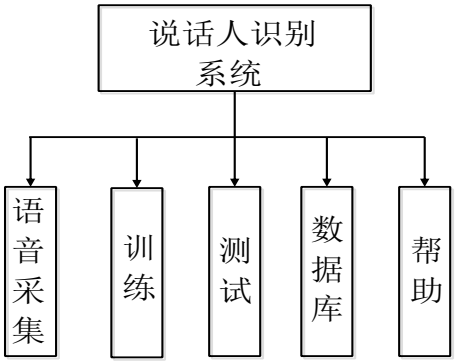



图 5-4 程序模块图

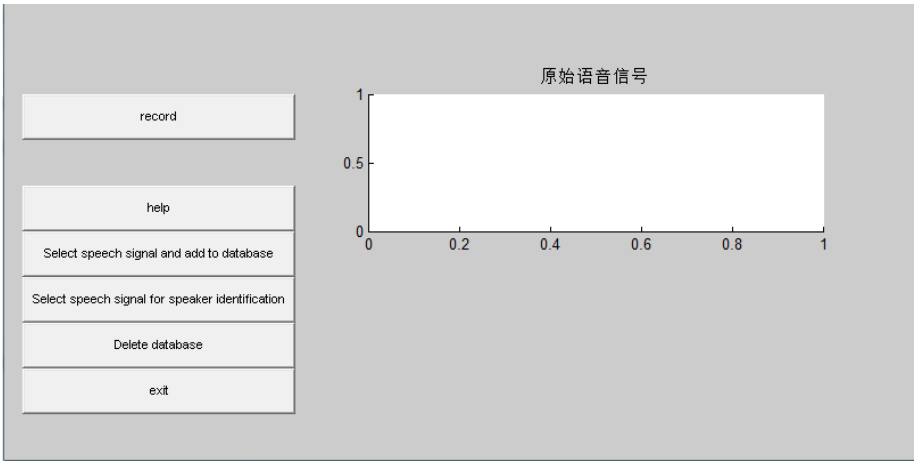



图 5-5 程序开始界面

```
%本程序为 matlab 的执行部分，执行该程序可完成说话人识别
set (gcf,'Position',[400,100,800,400])
bacg=figure(1)
%help 控件
h_push1=uicontrol(bacg,'style','push',...
        'unit','normalized','position',[0.02,0.5,0.3,0.1],...
        'string','help','callback','help');
%往数据库中添加语音特征向量
h_push2=uicontrol(bacg,'style','push',...
        'unit','normalized','position',[0.02,0.4,0.3,0.1],...
        'string','Select speech signal and add to database','callback','add_record');
%验证
h_push3=uicontrol(bacg,'style','push',...
        'unit','normalized','position',[0.02,0.3,0.3,0.1],…
```

```
        'string','Select speech signal for speaker
        identification','callback','identification');
%录音
h_push4=uicontrol(bacg,'style','push',...
        'unit','normalized','position',[0.02,0.7,0.3,0.1],...
        'string','record','callback','record');
%删除数据库中的某个数据
h_push5=uicontrol(bacg,'style','push',...
        'unit','normalized','position',[0.02,0.2,0.3,0.1],...
        'string','Delete database','callback','delete_data');
h_push6=uicontrol(bacg,'style','push',...
        'unit','normalized','position',[0.02,0.1,0.3,0.1],...
    'string','exit','callback','exit');    %退出
subplot(211);
set(gca,'Position',[0.4,0.5,0.5,0.3]);
title('原始语音信号');
```

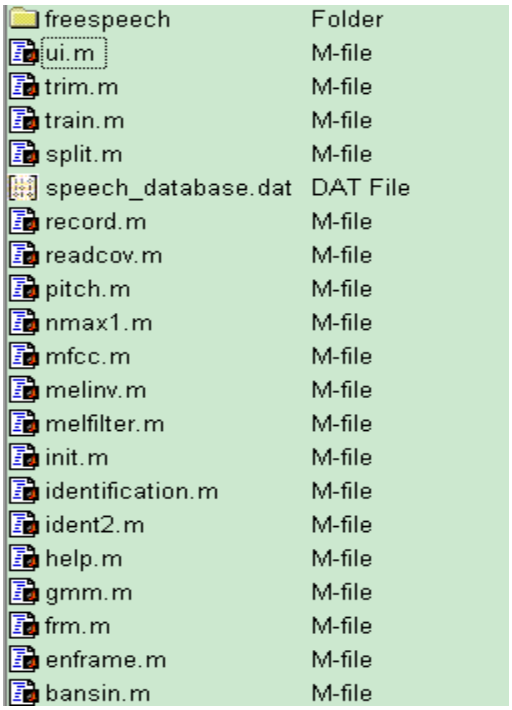



图 5-6 主要程序代码截图

图 5-6 为主要程序代码的截图，其中界面程序为 ui.m 主要代码上面已经给出，

主要的录音都在 freespeech 文件夹中。record.m 为录音程序，frm.m 包含了语音信号的去除寂静音，降噪，分帧等内容，pitch.m 为使用自相关法来求取基音周期的程序，mfcc.m 为 MFCC 特征向量的提取部分的程序，init.m 对高斯混合模型的均值，方差，以及系数进行了初始化，gmm.m 实现了期望值最大法来建立高斯混合模型，ident2.m 为利用最大似然概率来寻找输入数据的高斯模型。所有已经建模的数据都保存在 speech_database.dat 中，它包含了说话人的姓名、高斯混合模型的各个参量、以及基音周期等信息。

图 5-7 为录音开始时的提示，本程序默认录音时间为 20 秒，可自定义录音时间，具体方法为在 matlab 命令行输入形如 record('120s\','record10.wav')的命令即可，其中录音单位为秒。图 5-9 为一语音信号的 MFCC 特征向量图，其中特征向量的维数为 20 维（未加权前）。图 5-10 为模型训练过程的 matlab 命令行输出的数据截图，可见随着迭代次数的增加，log 似然值会不断的变大，直至达到设定的阈值或者迭代次数为止。

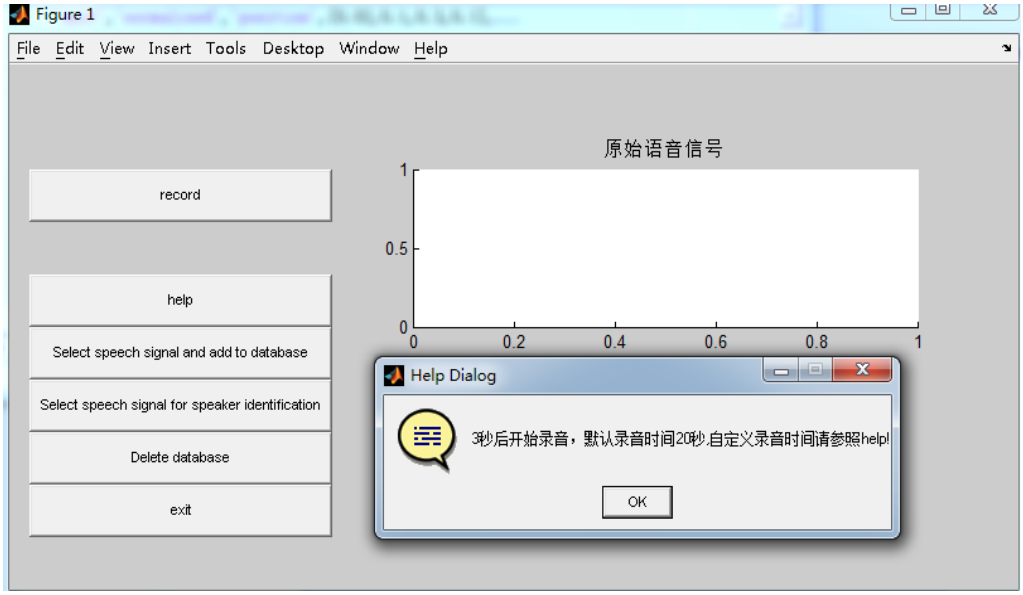



图 5-7 程序录音界面

图 5-8 为训练模块的流程图，主要对语音数据进行特征参数提取、训练，然后将得到的数据加入数据库中。该部分的主要代码如下：

```
[namefile,pathname]=uigetfile({'*.wav','speech Files (*.wav)'},'Chose speech signal');
fr1=frm(strcat(pathname,namefile),16,8000,1);%对语音信号进行分帧预加重处理等
subplot(212);
filt=melfilter(150,300,15);
```

```
v=train(fr1,filt,20);                    %获得 mfcc
imagesc(v);                              %显示 MFCC 特征向量的 RGB 图形
[a b]=size(v);
axis([1 b 1 a]);
ylabel('特征矢量维数');
xlabel('特征矢量个数');
colormap(hsv(255));
axis('xy');
data=struct('name',{},'means',{},'cov',{},'prob',{},'pitch',{});
prompt={'输入要添加的说话人姓名'};
name='the speaker ';
numlines=1;
defaultanswer={'no one'};
answer=inputdlg(prompt,name,numlines,defaultanswer);
data(speaker_number).name=answer{1,1};
data(speaker_number).means=nim1;
data(speaker_number).cov=readcov(nis1);
data(speaker_number).prob=nip1;
data(speaker_number).pitch=s1;
save('speech_database.dat','data','speaker_number');      %保存说话人的信息到数据库中
```

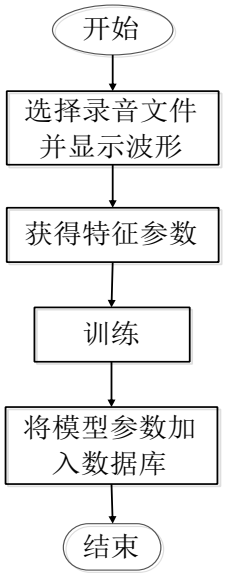



图 5-8 训练模块流程图

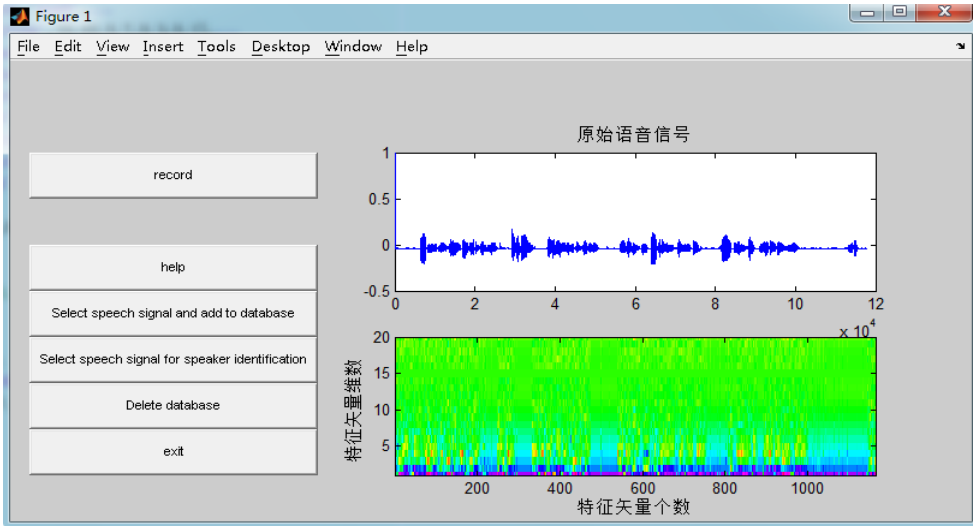



图 5-9 特征向量提取截图

图 5-11 为测试结果矩阵，我们取 3 个说话人的录音进行建模，然后在取这三个人的语音进行测试，矩阵的每一行为一个说话人作为训练模板，第 $(i, j)$ 个数为测试结果，如第 1 行第 2 列的数字含义为第一个说话人的语音做为训练语音，第二个说话人的语音做为测试语音，很明显对角线元素为正确的测试结果（在每一行中最大），观察矩阵也可以得出这一点，通常测试结果取阈值为-25，当所有的结果都小于-25 时可认为模板库中不存在该人。

```
log-likelihood :  -572.137892
log-likelihood :  -4.076585
log-likelihood :  -0.371515
log-likelihood :  -0.287362
log-likelihood :  -0.226704
log-likelihood :  -0.167614
log-likelihood :  -0.105148
log-likelihood :  -0.044949
log-likelihood :  0.007357
log-likelihood :  0.049708
       14499          12

Completed Training Speaker 1 model (Press any key to continue)
```

图 5-10 训练过程

```
log-likelihood :  -11.581529
Completed Training Speaker 3 model (Press any key to continue)
The comparisons Results
Note:
Each column i represents the test recording of Speaker i
Each row i represents the training recording of Speaker i
The diagonal elements (corresponding to same speaker comparisons
-------------------------------------------------------------------

A =

   -8.2118   -0.8261  -16.2001
   -9.6291    0.1869  -19.5029
  -29.5302  -45.5317  -11.6444

-------------------------------------------------------------------
>>
```

图 5-11 识别结果矩阵

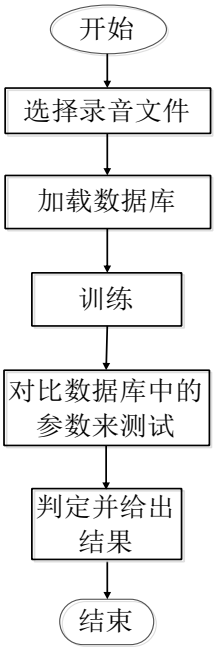



图 5-12 测试模块流程图

图 5-12 为测试模块的流程图，测试部分首先取一段测试的语音，加载已经训练的数据库，在对该段语音进行训练得到说话人语音的特征向量，通过测试与已知数据库中模型的相似度来判断是否为正确的说话人。

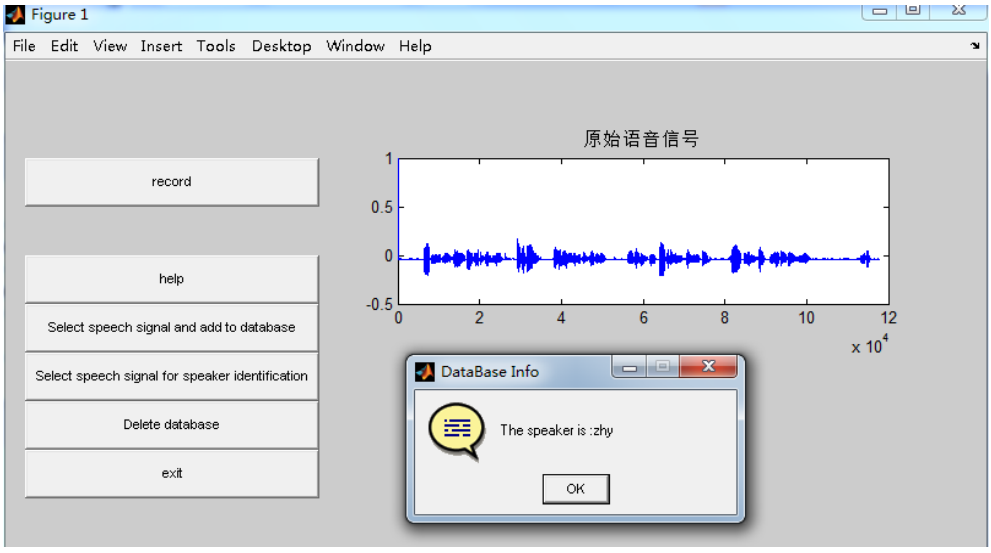



图 5-13 一次测试结果图

## 5.6 实验和结果分析

### 5.6.1 LPCC 和 MFCC 的实验比比较

在常用的特征参数中有线性预测倒谱系数 LPCC 和梅尔倒谱系数 MFCC，通常认为梅尔倒谱系数的性能更好，下面通过实验来验证。本实验取 20 人的语音，详细参数列于表 5-1。

表 5-1 特征向量对比实验参数设定

| | |
|---|---|
| 特征向量 | 16 维 MFCC，12 维 LPCC |
| 高斯分布个数 | 16 维 GMM |
| 训练语音长度 | 50s |
| 测试语音长度 | 3.2 秒，6.4 秒,9.6 秒,12.8 秒 |

表 5-2 特征向量对比实验结果

| 时长 | 3.2s | 6.4s | 9.6s | 12.8s | 16s |
|---|---|---|---|---|---|
| LPCC 识别率 | 75.66% | 82.14% | 84.39% | 85.81% | 87.44% |
| MFCC 识别率 | 78.17% | 86.12% | 87.73% | 87.80% | 89.62% |

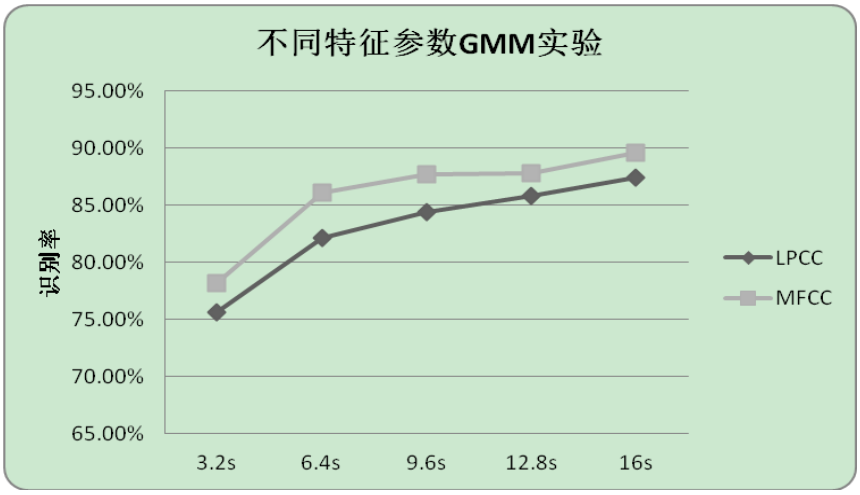



图 5-14 不同特征参数实验结果

从实验结果可以看出 MFCC 参数的识别效果总体上要好于 LPCC，这是因为在线性预测的时候声道模型被当作了全极点模型，而实际中声道响应中可能会包含零点，所以会造成结果不准确，现如今单纯的使用某一种特征向量已经很少，很多时候都是采用他们几个的结合的混合特征参数。

### 5.6.2 高斯混合模型阶数大小对识别的影响

基于高斯混合模型的说话人识别系统的高斯混合阶数会影响到识别率的高

低。若混合阶数太小则模型可能不能很好的表征说话人，从而引起识别率的下降，但是若阶数太大的话，则会增加训练和识别的时间。目前并没有理论可以用来估计合适的高斯分布个数。实验中所用的录音数据是 20 人的录音样本。详细参数列于下表。得到的识别率结果如表 5-3 所示。

表 5-3 不同阶数高斯分布实验参数

| 特征向量 | 16 维 MFCC |
|---|---|
| 高斯分布个数 | 4、16、32、64、128 |
| 训练语音长度 | 30s、60s、90s、120s |
| 测试语音长度 | 1s、3s、5s |

表 5-4 实验结果

| 训练材料量 | 测试长度 | 高斯分布个数 | | | | | |
|---|---|---|---|---|---|---|---|
| | | M=4 | M=8 | M=16 | M=32 | M=64 | M=128 |
| 30s | T=1s | 43.91 | 56.16 | 63.39 | 66.57 | 64.28 | |
| | T=3s | 60.81 | 72.73 | 77.15 | 78.25 | 73.68 | |
| | T=5s | 65.42 | 76.96 | 80.75 | 81.72 | 77.46 | |
| 60s | T=1s | 48.70 | 61.33 | 73.13 | 76.87 | 78.14 | 76.07 |
| | T=3s | 67.29 | 77.94 | 86.07 | 88.54 | 88.11 | 84.82 |
| | T=5s | 73.05 | 81.71 | 89.47 | 90.56 | 90.44 | 87.23 |
| 90s | T=1s | 50.61 | 64.37 | 77.23 | 82.93 | 84.63 | 84.57 |
| | T=3s | 69.56 | 81.23 | 89. 58 | 92.94 | 92.96 | 92.64 |
| | T=5s | 74.44 | 84. 94 | 93. 16 | 94.48 | 94.82 | 94.40 |
| 120s | T=1s | 53.13 | 66.91 | 79.73 | 85.06 | 86. 61 | 87.83 |
| | T=3s | 73.17 | 83.56 | 92.13 | 94.16 | 94. 18 | 94.4 |
| | T=5s | 78.33 | 87.29 | 94.71 | 95.95 | 96.03 | 95.84 |

表 5-4 为完整的实验结果，我们将部分的实验数据画成图 5-15 到图 5-18。图 5-15 及图 5-16 分别为固定训练时长为 60s 及 120s，不同的测试语句长度(1s,3s,5s)下高斯分布个数的比较图。图 5-17 和图 5-18 为分别为固定测试语句长度为 3s 和 5s 时不同训练材料下高斯分布个数的比较图。

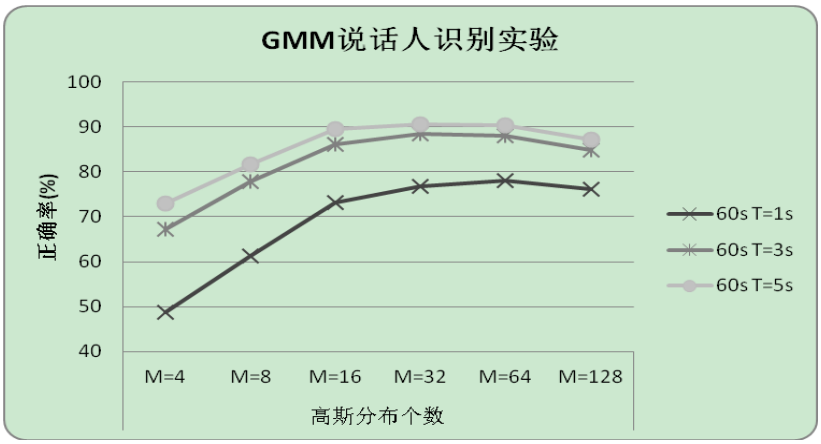



图 5-15 训练时间为 60s

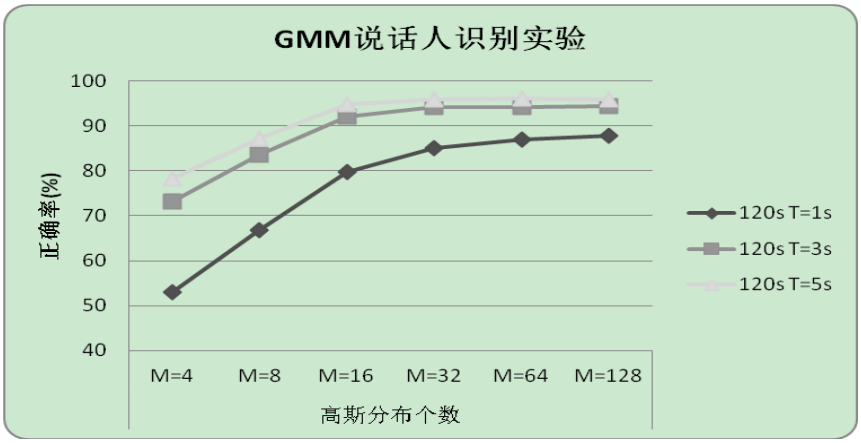



图 5-16 训练长度为 120 秒

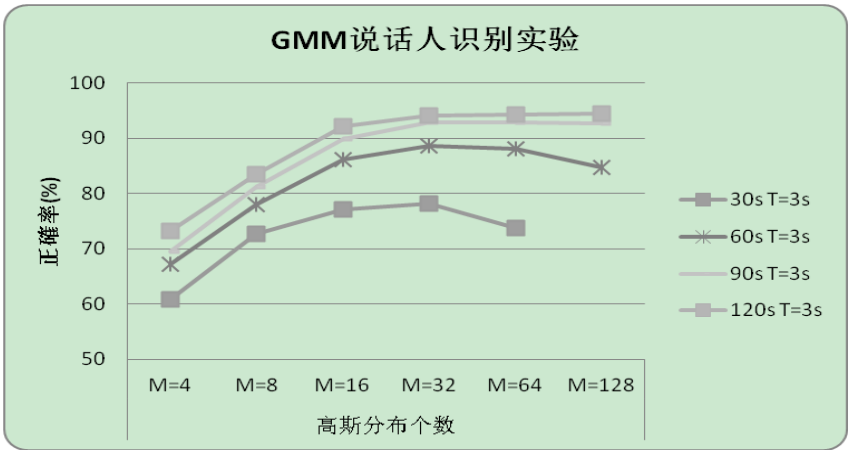



图 5-17 测试时间为 3s

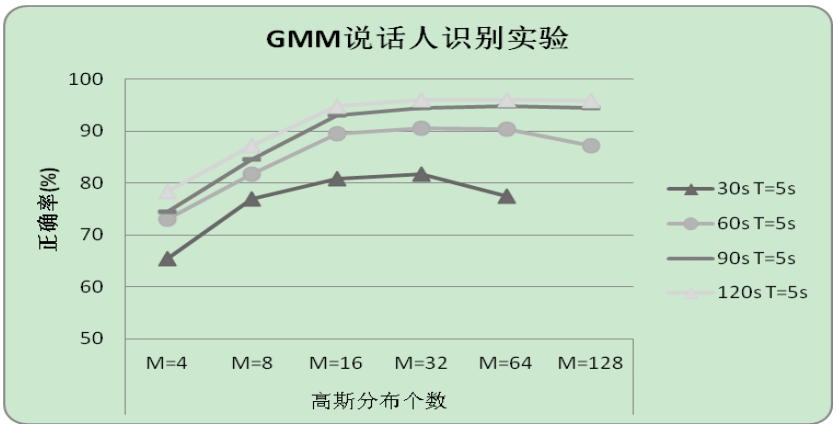



图 5-18 测试时间为 5s

我们由上面四个图可以发现，当高斯混合阶数小于 16 时，随着高斯混合阶数的增加识别率会随之增加。当高斯混合阶数为 16 以上以后，识别率的变化不明显，甚至识别率还会下降。比较图 5-15 及图 5-16，可以发现训练语音时长为 60s，识

别率从高斯混合阶数大于 64 以后开始下降，并且下降比较明显。当训练时间为 120 秒时，识别率在 128 以后有了小幅度下降。所以高斯混合阶数和训练语音的长度有关。当训练时间较长的时候，可以适当的增大混合阶数，如果训练时间较短的时候混合阶数过大反而会降低识别率。

另外一个是测试语句对识别率影响的讨论，毫无意外的，由四个图的表现来看，测试语句越长对辨识率的提升越有帮助。而且比较图 5-17 及 5-18，我们可以发现测试语句越长（5s）时，各种训练语料量所得出辨识结果的差距会比测试语句（3s）短时小。本文取 GMM 阶数为 16 维。

### 5.6.3 MFCC 维数对识别结果的影响

在第二章信号的倒谱分析部分已经介绍过人的声道激励特性对应于倒谱参数的低维部分，本实验设定高斯分布个数为 16 维的情况下对 MFCC 参数的阶数不同的情况下识别率进行试验，试验中取 20 人的录音，训练时间为 20 秒，测试时间为 1s。测试结果如表所示。

表 5-5 不同阶数的 MFCC 实验参数

| MFCC 阶数 | 5 | 7 | 9 | 12 | 14 | 16 | 18 | 20 | 25 |
|---|---|---|---|---|---|---|---|---|---|
| 识别率(%) | 37.91 | 47.27 | 58.86 | 55.04 | 58.18 | 58.14 | 56.13 | 55.92 | 52.95 |

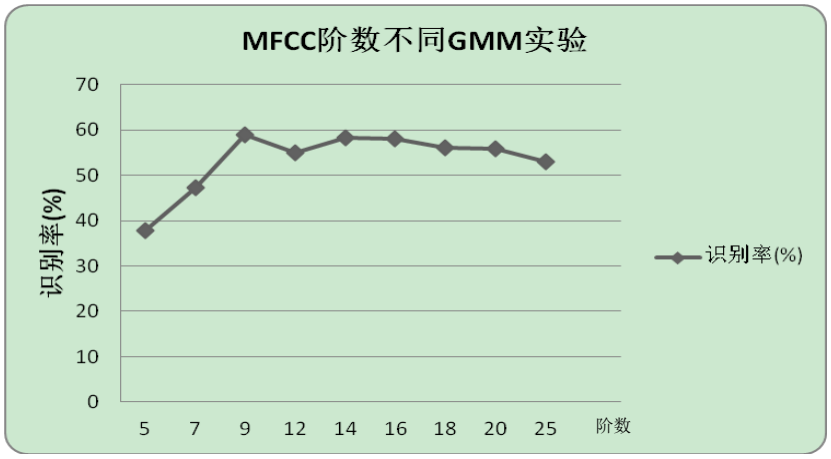



图 5-19 不同特征参数实验结果

从实验结果可以看出在 MFCC 阶数高于 16 阶以后，说话人识别率并没有相应的提高，而且可能造成识别效率的下降，这是因为随着阶数的增加 MFCC 参数中包含无效的语音特征信息，从而造成了识别结果的下降，这与倒谱分析部分所介绍的情况相符，本文 MFCC 的阶数取 16。

### 5.6.4 实验加入动态 MFCC

在第三章 MFCC 特征参数提取过程中已经介绍过，对于语音信号都是假设为短时平稳的，所以在信号处理的过程中都要进行加窗，这样就造成了信号的不连续性，从而导致了语音动态部分的丢失。本实验通过取 MFCC 参数和 MFCC 参数的一阶差分进行识别率对比。实验中所用的样本为 20 人的录音。详细设置如表 5-6 所示。

表 5-6 不同的 MFCC 实验参数

| | |
|---|---|
| 特征向量 | 16 维 MFCC， MFCC+$\Delta$MFCC+$\Delta^2$MFCC（共 48 维），16 维$\Delta$MFCC |
| 高斯分布个数 | 16 |
| 训练语音长度 | 50 秒 |
| 测试语音长度 | 3.2 秒、6.4 秒、9.6 秒，12.8 秒，16 秒 |

表 5-7 不同的 MFCC 实验结果

| 测试时间 | 3.2s | 6.4s | 9.6s | 12.8s | 16s |
|---|---|---|---|---|---|
| MFCC | 75.47% | 83.01% | 85.74% | 86.91% | 88.62% |
| ∆ MFCC | 50.31% | 56.36% | 64.12% | 65.37% | 71.75% |
| MFCC+∆ MFCC+∆ 2MFCC | 83.17% | 88.15% | 91.37% | 91.93% | 92.75% |

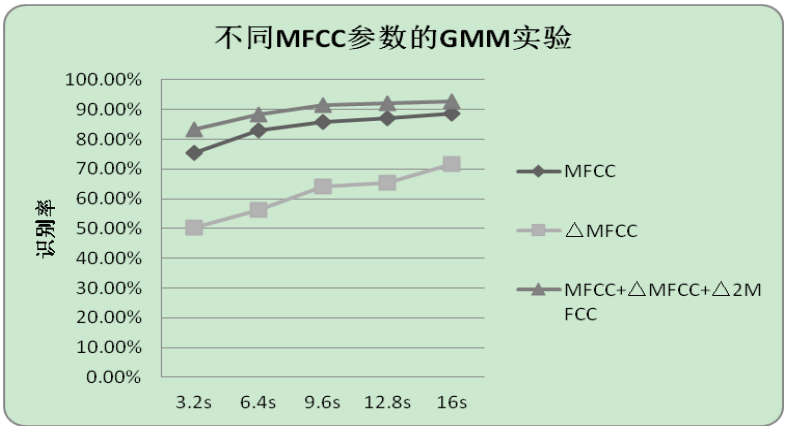



图 5-20 不同 MFCC 特征参数实验结果

通过实验可以看出一阶差分 MFCC 的识别率最低，加入动态 MFCC 参数后识别率最高，单独使用一阶 MFCC 虽然识别效果很差，但是动态 MFCC 中依然包含了有用的说话人特征信息，通常在实际应用中和 MFCC 结合，加入动态 MFCC 的识别率有所提高，但是这是以牺牲运算复杂度为代价的（由原来的 16 维变成 48 维）。

### 5.6.5. 改进的加权 WMFCC 实验

通过上面的实验可以得知加入动态的 MFCC 可以提高识别率，但是运算量会大大增加，基于这个原因本文使用了加权的 MFCC 称为 WMFCC，根据对人声音特征的贡献大小通常一阶二阶差分 MFCC 的系数要小于 1，根据经验设置一阶 MFCC 的系数为 1/3，二阶 MFCC 的系数设置为 1/6，这样可以达到降维的目的，同时提高了识别率。实验中所用的样本为 20 人的录音。详细设置如表 5-8 所示。

表 5-8 加权的 MFCC 实验参数设定

| | |
|---|---|
| 特征向量 | 16 维 MFCC ， MFCC+$\Delta$MFCC+$\Delta^2$MFCC（共 48 维），16 维 WMFCC(一阶 MFCC 系数为 1/3，二阶 MFCC 系数为 1/6) |
| 高斯分布个数 | 16 |
| 训练语音长度 | 50s |
| 测试语音长度 | 3.2s、6.4s、9.6s、12.8s、16s |

表 5-9 加权的 MFCC 对比实验结果

| 测试时间 | 3.2s | 6.4s | 9.6s | 12.8s | 16s |
|---|---|---|---|---|---|
| MFCC | 75.47% | 83.01% | 85.74% | 86.91% | 87.35% |
| WMFCC | 80.27% | 86.36% | 87.52% | 90.37% | 92.18% |
| MFCC+△ MFCC+△ 2MFCC | 83.17% | 88.15% | 91.37% | 91.93% | 92.75% |

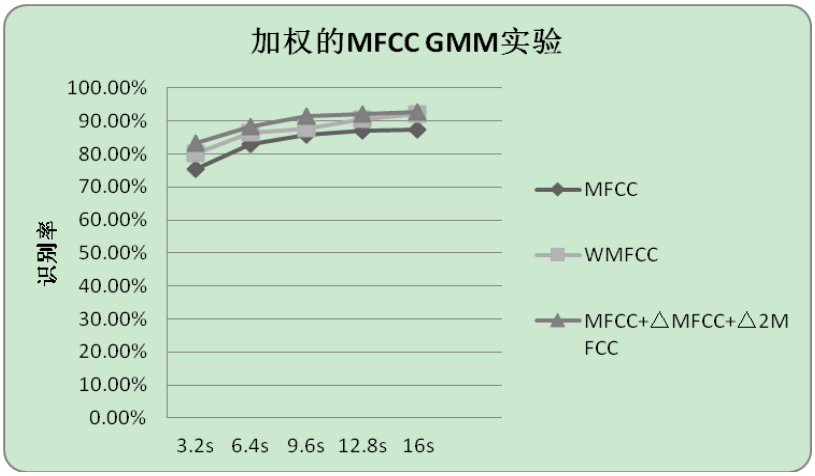



图 5-21 加权的 MFCC 对比的实验结果

通过实验结果可知加权后的 MFCC 识别效果与加入动态 MFCC 的识别效果相差不大，同时又高于一般的 MFCC 的识别率，这就为识别效果和系统的运算复杂度之间找到了一个平衡，通过实验验证，加权后的 MFCC 简单有效。

### 5. 6. 6. 加入基音周期的实验

前面章节中只使用了 16 维的特征参数 MFCC 进行识别，MFCC 包含语音频率结构的时间变化信息，反应了声道运动的动态特征，也就是发音方式、发音习惯等，相对稳定，但是比较容易模仿，针对这种情况，我们在特征维中加入基音周期（pitch），基音周期包含了语音频率结构信息，反应了声带的特征，容易受健康状况的影响，但不容易模仿。这样，特征数据即为 17 维的数据，其中，前 16 维为 16 的 MFCC，第 17 维为 pitch。实验中取 20 人的录音做为样本。详细设置如表 5-10 所示。

表 5-10 加入基音周期的实验参数

| | |
|---|---|
| 特征向量 | 16 维 MFCC ，16 维 MFCC+pitch（共 17 维） |
| 高斯分布个数 | 16 |
| 训练语音长度 | 50s |
| 测试语音长度 | 3.2s、6.4s、9.6s、12.8s、16s |

表 5-11 加入基音周期的实验结果

| 测试时间 | 3.2s | 6.4s | 9.6s | 12.8s | 16s |
|---|---|---|---|---|---|
| MFCC | 76.13% | 83.21% | 85.84% | 87.01% | 88.16% |
| MFCC+pitch | 75.27% | 85.29% | 86.52% | 89.33% | 91.27% |

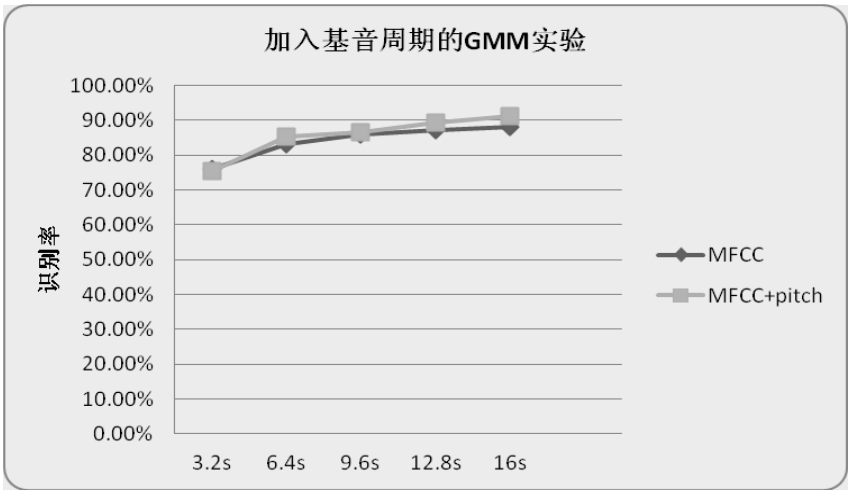



图 5-22 加入基音周期的 GMM 实验

从实验结果可以看出加入基音周期后的 MFCC 参数识别效果和 MFCC 近似，在测试时间较短的时候（3.2s）甚至低于 MFCC 的识别效果。虽然 pitch 可以反映人的独有特征信息，但是由于人的生理状况的不同可能会随时改变，所以在实际应用中，并不常见，但是做为对声音信息的一个很好的表征，在说话人识别的另一个分支说话人确认（speaker verification）可能会更多的用到，这是一个值得我们

继续研究的课题。

## 5.7 基于改进的帧投票判决方法

由于语音中含有无效的语音帧，根据说话人辨认的判决等式可知，传统的判决方法中若无效语音帧的数目比较多，则会对判决结果造成很大的影响。本文使用帧投票的方法做实验如下。实验条件如下:样本人数为 20 人的录音。详细设置如表 5-12 所示。

表 5-12 不同的判定方法试验参数

| | |
|---|---|
| 特征向量 | 16 维 MFCC |
| 高斯分布个数 | 16 |
| 训练语音长度 | 50s |
| 测试语音长度 | 3.2s、6.4s、9.6s、12.8s、16s |

表 5-13 不同识别统计方法得到的识别率结果

| 测试时间 | 3.2s | 6.4s | 9.6s | 12.8s | 16s |
|---|---|---|---|---|---|
| 传统判决方法 | 76.46% | 83.31% | 85.29% | 87.46% | 88.16% |
| 帧投票判决方法 | 78.27% | 86.98% | 87.52% | 90.33% | 93.27% |

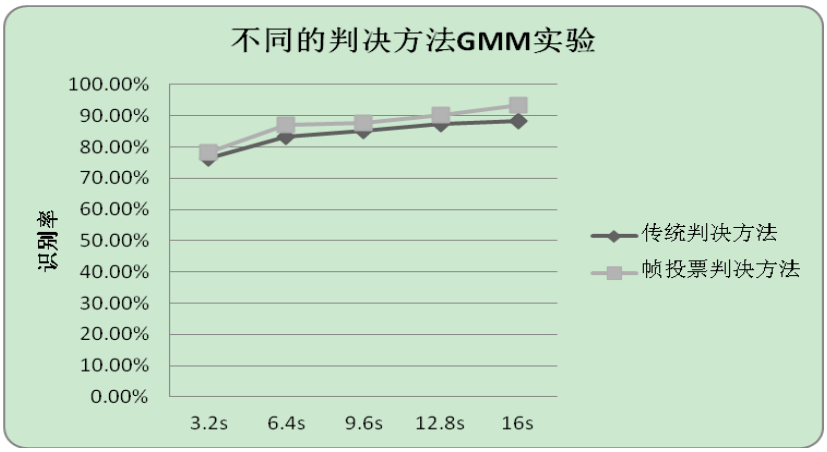



图 5-23 不同识别统计方法得到的识别结果

通过实验可以看出基于帧投票的判决方法比传统的判决方法识别效果更好，特别是测试时间比较长的时候，这是因为随着测试时间的加长，总的语音帧中的非正常干扰帧会增多，从而对于测试的整体性能影响比较大。由于实验条件所限，本文所用的语音库人数比较少，没有出现投票法的效果严重下降的情况。

# 第六章 总结和展望

## 6.1 总结

近年来随着语音处理技术的发展，利用语音作为识别工具的应用越来越多，本文讨论了基于高斯混合模型的说话人识别系统，该系统的一个主要特点就是文本无关性。本文详细分析了梅尔倒谱系数的提取，以及高斯混合模型的建立，识别过程。

本文针对梅尔倒谱系数的不足，给出了几点改进：在加窗环节采用更好频谱特性的凯泽窗来代替传统的汉明窗，并在 FFT 计算频谱的时候采用绝对值，这样可以减少计算量。针对 MFCC 会丢失动态信息的问题，传统的方法是加入 MFCC 的差分信息来解决，但是这样就会增加特征向量的维数，本文采样了加权的方法根据实验验证，该方法有着良好的效果。由于 MFCC 虽然相对稳定，但不同的说话人之间容易相互模仿，本论文针对 MFCC 的易模仿性，增加基音周期特征参数，基因周期包含了语音频率结构信息，虽然会受到说话人健康状况的影响，但不容易模仿。本文将两者相结合来应用到说话人识别中去。

说话人识别的第二部分就是对每个说话人的提取后的特征向量进行建模，在识别阶段传统的方法因为存在部分帧得分太低从而会影响整个评判结果，本文采用了帧投票的方法来进行解决。为了应对冒充者的问题，我们提取了基音周期并在测试部分加入到提取的 MFCC 特征向量中，针对 VQ 和 GMM 的各自的优缺点本文将它们相结合，使用了 VQ/GMM 模型，该模型首先将说话人的 GMM 模型分成了男声和女声两类，然后在使用 GMM 来识别，并使用 DTW 算法来优化识别过程从而减少了计算量，节省了识别时间。

## 6.2 未来展望

由于条件限制，本文中测试使用的采用的语音库比较小，可以使用国外通用的大型语音库来对实验结果进行验证。根据说话人识别的特点下一步的工作准备按照以下几个方面来进行：

（1）提高系统的抗噪性能，比如现有的解决方法中已经存在的向量自回归高

斯混合模型。

（2）寻找更好的特征参数，包括运算复杂度，是否能更好的表征人的声音特征等几个方面。

（3）在建模的阶段，可以加入非线性部分

（4）模型中的非正交化问题。

（5）环境匹配问题，如麦克风、录音环境等的不同。

（6）实现是在 PC 上使用 Matlab 编程工具来实现，由于嵌入式系统的应用越来越广泛，下一步可以将说话人识别系统移植到嵌入式设备中去。

（7）如今的研究的特征参数主要集中在低级别部分（LPCC，MFCC），可将人语音中的高级别部分加入特征向量中去[52]（语言的语义内容，即语言所要表达的意思），使得识别更加精确。

（8）本文使用的录音数据是在实验室条件下进行录音，但是录音数据量比较小，由于国内的现有中文录音数据库比较少，没有一个权威的中文数据库，下一步可采用国际上通用的英文数据库，或者增加录音人数来进行训练测试。

# 致　谢

时光匆匆，转眼间研究生生活就将步入尾声，这三年经历了很多，这篇论文的完成要感谢很多人，首先要感谢我的导师李绍荣教授，从论文的选题到研究方法李老师都给予了我精心的指导，最重要的是李老师以他严谨的治学学风，踏实的科研作风，实事求是的处理问题方法，对我今后的人生都将有着深远的影响，必将使我受益终生。在此献上我最诚挚的感谢以及衷心的祝福。

其次要感谢教研室的同学，不论在生活上还是学习上，他们都给与了我很多帮助，他们帮助使我有勇气面对各种困难，并在论文的撰写过程中给我提出了各种宝贵的意见使我能够顺利完成论文。

最后还要感谢我的家人，他们的关怀一直支持着我努力向前。在我面临各种压力的时候一句鼓励的话就使我有了无尽的动力！

# 参考文献

[1] 郭晓丹. 计算机语音识别技术问题漫谈. 自动化信息. 2005, 49(5):43-45

[2] 黄颖. 基于 GMM 的与文本无关的变阈值说话人确认. 成都信息工程学院学报, 2004, 19(4):541-544

[3] 邓浩江. 基于聚类统计与文本无关的说话人识别研究. 电路与系统学报, 2001, 6(3):76-80

[4] 蒋晔. 基于文本无关的说话人识别技术研究:[硕士学位论文] , 南京：南京理工大学, 2008, 3-6

[5] D.A. Reynolds.Speaker identification and verification using Gaussian mixture speaker models.Speech Communication,1995,60(17):91-108

[6] ChiWei Che ,CAIP Center, Rutgers Univ. An HMM approach to text independent speaker verification.In IEEE International conference on Acoustics,1996,2:673-676

[7] W.M.Campbell, J.P. Campbell, D.A. Reynolds, E. Singer. Torres-Carrasquillo, Support vector machines for speaker and language recognition.Computer Speech and Language,2006,70(2):210-29

[8] F.K. Soong.A vector quantization approach to speaker recognition, AT & T Technical Journal,2004 100(66):14-26

[9] S. Bengio, J. Mariethoz.Learning the decision function for speaker verification.In IEEE International conference on Acoustics,2001,12(1):425-428

[10] Xiaoyuan Zhu, Bruce Millar,Iain Macleod,Michael Wagner.A comparative study of mixture-Gaussian VQ,ergodic HMMs and left-to-right HMMs for speaker recognition.In international symposium on Speech,Image processing and Neural Networks, 1994,70(3)：618-621.

[11] S.Furui.Vector quantization based speech recognition and speaker recognition techniques.systems and computers,1991,40(2):954-958

[12] Lawrence R.Rabiner,Ronald W. Schafer.Introduction to digital speech processing Now Publishers Inc,2007,55-67

[13] 俞一彪. 说话人语音特征子空间分离及识别应用.电路与系统学报. 2008, 13(1):158-163

[14] R.A. Finan, A.T. Sapeluk, R.I Damper.Comparison of multilayer and radial basis function neural networks for text independent speaker recognition.In IEEE international conference on

neural networks.1996,40(23):133-139

[15] Price,Willmore,Roberts.Genetically Optimized Feed forward Neural Networks for Speaker Identification.Information Technology Division, Electronics and Surveillance Research Laboratory.2008,40-52

[16] 段盼爽. 人工语音带宽扩展算法研究:[硕士学位论文]. 大连：大连理工大学, 2008, 20-30

[17] 林焕祥. 言语过滤识别.远程教育杂志, 2004, 163(4):48-50

[18] P.Rose.Forensic Speaker Identification.Taylor&Francis,London.2002,174(2):80-83

[19] 李泽.MFCC 和 LPCC 特征参数在说话人识别中的研究. 河南工程学院学报, 2010, 22(2):52-55

[20] 汤小飞. 基于全局背景模型和辅助模型的说话人确认系统：[硕士学位论文]. 南京：南京师范大学, 2010, 30-32

[21] 韩纪庆, 张磊, 郑铁然. 语音信号处理. 北京：清华大学出版社出版社, 2004, 10-50

[22] 张云雁. 一种基于改进的矢量量化算法的说话人识别方法. 上海大学学报, 2005, 11(4):368-371

[23] 刘雪燕. 说话人识别综述. 电脑知识与技术, 2009, 5(1):81-90

[24] H.Hermansky,N.Morgan,A.Bayya,and P.Kohn.Compensation for the effect of thecommunication channel in auditory-like analysis of speech (rasta-plp).Proceedings European Conf.on Speech Communication and Technology.EUROSPEECH, 1991:1367-1370.

[25] 易克初. 语音信号处理. 北京：国防工业出版社, 2000, 10-70

[26] 胡国强. 基于线性预测残差倒谱的多语音基音频率检测. 电子技术, 2009, 36(12):15-18

[27] 石海燕. 语音信号特征参数研究.电脑知识与技术, 2008, 17(3):754-757

[28] 徐正伟. 语音信号处理及其在 IP 网络电话中的应用. 今日电子, 2001, 93(11):20-21

[29] Zheng-Hua Tan, Borge Lindberg Automatic speech recognition on mobile devices and over communication networks Springer,2008,40(3):68-83

[30] Ben Gold, Nelson Morgan, Dan Ellis .Speech and Audio Signal Processing: Processing and Perception of Speech and Music.Technology & Engineering,2011,40(8):440-470

[31] 李永宁. 基于自相关的语音基音周期检测方法研究.福建电脑, 2008, 24(11):59-62

[32] 杨森斌. 一种改进的自相关函数基音检测算法. 现代电子技术, 2008, 272(9):135-137

[33] 杨莎. 一种基音周期估计的改进 CAMDF 算法. 四川大学学报, 2008, 45(4):773-778

[34] Hwai-Tsu,Hu Chu,Yu:Chih-Hang Lin.Usefulness of the Comb Filtering Output for Voiced/Unvoiced Classification and Pitch Detection .International Conference on Signal Processing Systems,2009,78(4):63-65

[35] 孙燕. 声门激励信号的获取及其应用.电脑开发与应用, 2010, 23(8):13-15

[36] 潘秀林. 一种有效的清浊判决及其基音周期估计方法. 江苏航空, 2007, 147(4):13-15

[37] Lasse L Molga ,W Jorgensen .Speaker Recognition:Special Course,2005,93(2):15-22

[38] 张奇. 基于支持向量机的乐器识别方法. 计算机工程与应用, 2004, 83(18):99-101

[39] 许庆晗. 基于语音和人脸的身份认证技术研究:[硕士学位论文]. 南京：东南大学，2009, 20-31

[40] Patel.Modified MFCC Windowed Technique for Speaker Word Recognition.International Conference and Workshop on Emerging Trends in Technology,2011,196(3):1311-1315

[41] Politeknik Seberang Perai.Robust Computer Voice Recognition Using Improved MFCC Algorithm.International Conference on New Trends in Information and Service Science, 2009,432(4):70-76

[42] 甄斌. 语音识别和说话人识别中各倒谱分量的相对重要性. 北京大学报，2001, 37(3):371-378.

[43] Jorge S. Marques, Nicolas Perez De La Blanca, Pedro Pina ,Pattern Recognition and Image Analysis: Second Iberian Conference,2005,89(5):587-592

[44] 王松. 基于 GMM 与改进 MCE 训练的说话人识别研究. 自动化与仪器仪表，2010, 152(6):21-24

[45] 陈立伟. 一种基于混合神经网络说话人识别系统. 哈尔滨工程大学学报，2005, 26(6):781-784

[46] A.P.Dempster,N.M.Laird,and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B.,1997,39(1):30-38

[47] T Kinnunen Elsevier. An overview of text-independent speaker recognition: From features to supervectors-Speech Communication,2010,97(3):56-61

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

[49] P Kenny,G Boulianne,P Ouellet.Speaker and session variability in GMM-based speaker verification,IEEE,2007,1-14

[50] B.Narayanaswamy and Rashmi Gangadharaiah.Extracting additional information from gaussian mixture probabilities for improved text-independent speaker identification. In Proc.of IEEE,ICASSP,2005,621-624

[51] Balakrishnan.Improved Text-Independent Speaker Recognition using Gaussian Mixture Probabilities.Carnegie Mellon University,2005,10-12

[52] Sandipan Chakroborty,Goutam Saha.Improved Text-Independent Speaker entification using Fused MFCC & IMFCC .Feature Sets based on Gaussian Filter International Journal of Information and Communication Engineering,2009,5(1):11-19

[53]采用 DTW 算法和语音增强的嵌入式声纹识别系统，2012,51(2):174-178

[54] Code Separated Text Indentpent Speaker Identification System Using GMM,International Journal of Multimedia and Ubiquitous Engineering,2011,6(3):61-73

# 攻读硕士学位期间取得的研究成果

[1] 刘士.Linux 平台下的 ALSA 声音编程.计算机光盘软件与应用,2011(6):175-176