西南交通大学

硕士学位论文

说话人识别的研究与DSP实现

姓名：曾海涛

申请学位级别：硕士

专业：密码学

指导教师：史燕

20060401

摘 要

说话人识别又称为声纹识别，是指根据说话人的声音特征来识别说话人的技术，近年来，在生物识别技术领域中，声纹识别技术以其独特的方便性、经济性和准确性等优势受到瞩目，在信息安全领域的应用逐渐增大，并将日益成为人们日常生活和工作中重要且普及的安全验证方式。

本文回顾了说话人识别技术发展的历史，阐述了特征提取、模式匹配和判决规则等说话人识别中的主要技术理论。详细地讨论了基音频率、线性预测系数及美尔倒谱系数等特征提取方法，以及动态时间规整、矢量量化、隐马尔可夫模型等模式匹配算法的原理及实现流程。

本文的目标是要设计一个基于 DSP 的说话人识别系统，性能和复杂度是比较关键的因素。因此在对几种不同的算法方案进行分析比较的基础上，采用美尔倒谱系数作为特征参数，以矢量量化作为匹配算法设计了一个说话人识别系统，并对系统在不同参数下的识别性能进行了仿真，选取最优的方案在 TMS320 C5402DSK 上实现了该系统。论文的主要研究方向是针对系统的实现平台上的实际应用进行系统设计、提高系统的识别率、可靠性和减少识别时间。经过测试，系统运行正确，达到预期目标。

【关键词】：说话人识别；DSP ；矢量量化

Abstract

Speaker recognition, which is also called voice print, is a technology that attempts to recognize a speaker through measurements of the specifically individual characteristics arising in speaker's voice. Recent years, in the biological identification technologies, speaker recognition technology with unique advantages such as convenience, economy and accuracy is paid much attention to. In the field of information security, applications using speaker recognition technology is gradually increasing, and it will becoming an important and popular security certification method of daily life and work.

The thesis first briefly reviews the history of speaker recognition_and then discusses the main theory of it. Feature extraction model, pattern matching model and decision model as the key factors of speaker recognition, several models and algorithm for each part are discussed in detail.

The aim of this thesis is to design a practical speaker recognition system based on DSP, in which performance and complexity is relatively the key factors. Based on this conception, according to the compare of time and space consumption for each algorithm, MFCC and VQ are finally chosen to be implemented. Then the recognition performance on various system parameters is emulated to achieve best scheme. Finally the scheme is implemented on the platform of TMS320 C5402DSK system. The main effort then is paid to enhance recognition rate and reduce time and space consume. After testing, system operates correctly, and the target is reached.

Key Words: Speaker Recognition; DSP; Vector Quantization

第一章 绪论

1.1 说话人识别的概念

说话人识别又称为声纹识别，近年来，声纹识别技术以其独特的方便性、经济性和准确性等优势受到瞩目，在信息安全等领域的应用逐渐增大，并将日益成为人们日常生活和工作中重要且普及的安全验证方式。

声纹识别属于生物识别技术的一种，是一种根据语音波形中反映说话人生理和行为特征的语音参数，自动识别说话人身份的技术。与语音识别不同的是，声纹识别利用的是语音信号中的说话人信息，而不考虑语音中的字词意思，它强调说话人的个性，而语音识别的目的是识别出语音信号中的言语内容，并不考虑说话人是谁，它强调共性。

声纹识别系统主要包括两部分，即特征提取和模式匹配。特征提取的任务是选取唯一表现说话人身份的有效且稳定可靠的特征，模式匹配的任务是对训练和识别时的特征模式做相似性匹配。

与其他生物识别技术，诸如指纹识别、掌形识别、虹膜识别等相比较，声纹识别除具有不会遗失和忘记、不需记忆、使用方便等优点外，还具有以下特性：

- ◆ 用户接受程度高，由于不涉及隐私问题，用户无任何心理障碍。
- ◆ 利用语音进行身份识别可能是最自然和最经济的方法之一。声音输入设备造价低廉，甚至无费用，而其他生物识别技术的输入设备往往造价昂贵。
- ◆ 在基于电信网络的身份识别应用中，如电话银行、电话炒股、电子购物等，与其他生物识别技术相比，声纹识别更为擅长，得天独厚。
- ◆ 由于与其他生物识别技术相比，声纹识别具有更为简便、准确、经济及可扩展性良好等众多优势，可广泛应用于安全验证、控制等各方面，特别是基于电信网络的身份识别。

1.2 本论文的主要工作

本文的主要内容是研究说话人识别系统，并在 DSP 上实现，主要做了以下

几方面的工作：

1.研究了说话人识别的基础理论，对目前常用的几种特征提取和模式匹配算法进行了分析，对 LPC、MFCC 和 VQ、HMM 等算法的复杂性和效率进行了比较。

2.使用 Matlab 和 Visual C＋＋对几种算法进行了仿真，比较了算法在不同参数下的识别性能。

3.采用特征提取算法 MFCC 和模式匹配算法 VQ 的组合，在 TMS320 C5402 开发板系统上实现了说话人辨认系统。

4.优化了 DSP 算法的逻辑结构，使得系统的存储空间利用率和时间效率得以提高。

第二章 说话人识别理论

2.1 语音信号产生模型

在研究了发声器官和语音的产生过程以后，便可以建立一个离散时域的语音信号产生模型，对于进一步的各项研究以及各种具体应用，这个模型是非常重要的。这里先给出一个较简单的模型，(对于大多数研究和应用而言)，这个模型可以完全满足需要。

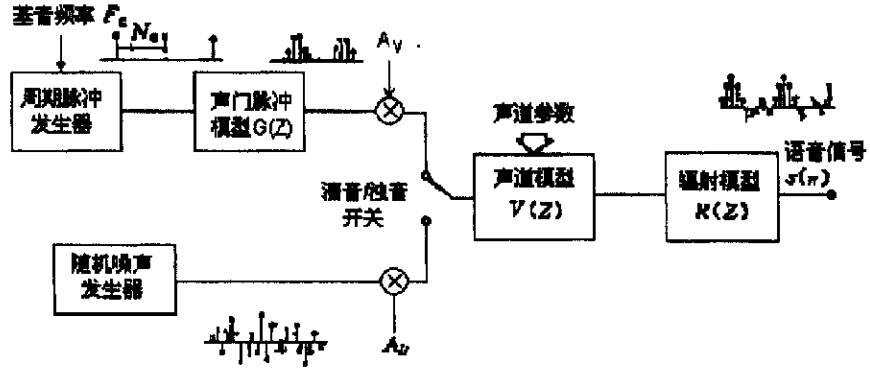


图 2.1 语音的离散时域模型

语音信号在较短时间可以认为是平稳的，在此前提之下，经典的语音信号模型可以用线性时不变系统来表示。为了了解语音信号的特性，给出语音产生的模型，如图 2.1 所示。研究表明，此模型可以满足大多数语音信号的研究和应用。

根据图 2.1 可以看出，语音信号离散时域模型包括三部分：激励源、声道模型和辐射模型。激励源部分是由浊音/清音开关所处的位置来决定产生的语音信号是清音还是浊音。

发出不同的音，激励的情况不同，大致分为两类：发浊音时，气流通过绷紧的声带，冲击声带产生振动，使声门处产生准周期性的脉冲串，脉冲串激励声道；发清音时，声带不振动，气流通过声门，类似于随机白噪声直接进入声道。因此，语音信号可以看成是在准周期脉冲或随机噪声激励下的输出。

声道模型 $V(Z)$给出了离散时域的声道传输函数，实际的声道被假设成一个

声道模型 $V(Z)$给出了离散时域的声道传输函数，实际的声道被假设成一个

变截面积的声管。采用流体力学的方法可以导出，在大多数情况下，声道模型是一个全极点函数。因此，*V(Z)*可以表示为：

$$V(Z)=\frac{1}{1-\sum_{k=1}^{p}a_k z^{-k}} \tag{2.1}$$

可以看出，这里把截面积连续变化的声管近似为 *p* 段短声管的串联。*p* 称为这个全极点滤波器的阶数。显然，*p* 值越大模型的传递函数与声道实际传输函数越吻合，一般 *p* 值取 8~12 即可。若 *p* 取偶数，*V(Z)*一般有 *p/2* 对共轭极点，分别与语音的各个共振峰对应。

辐射模型 *R(Z)*与嘴形有关，研究表明，口唇辐射在高频端较为显著，在低频端时影响较小，所以辐射模型 *R(Z)*为一阶高通滤波器。其表达式为：

$$R(Z)=R_0(1-z^{-1}) \tag{2.2}$$

在以上的模型中，*G(Z)* 和 *R(Z)*保持不变，基音周期、Av、Au、浊清音开关的位置以及声道模型参数中的参数 $a_1 \sim a_p$ 都是随时间变化的。考虑语音信号的短时性[11]，一般在语音信号分析中，取语音的分析帧长为 20ms 左右。

2.2 语音的特征提取

说话人识别系统中的特征提取即提取语音信号中表征个人的基本特征，此特征应能有效地区分不同的说话人，且对同一说话人的变化保持相对稳定。考虑到特征的可量化性、训练样本的数量和系统性能的评价问题，目前的说话人识别系统主要依靠较低层次的声学特征进行识别。说话人特征大体可归为下述几类：

谱包络参数，语音信息通过滤波器组输出，以合适的速率对滤波器输出抽样，并将它们作为声纹识别特征。

基音轮廓、共振峰频率带宽及其轨迹这类基于发声器官，如声门、声道和鼻腔的生理结构而提取的特征参数。

线性预测系数，如线性预测系数、自相关系数、反射系数、对数面积比、线性预测残差及其组合等参数，作为识别特征可以得到较好的效果，主要原因是线性预测与声道参数模型是相符合的。

反映听觉特性的参数模拟人耳对声音频率感知的特性而提出的多种参数，

如美尔倒谱系数、线性预测系数等。

2.2.1 基音周期

在人的发音模型中，产生浊音的周期激励脉冲的周期称为基音周期(Pitch)。只有浊音才有基音周期，而清音没有基音周期。基音周期或基音频率是语音信号的一个重要的参数。到目前为止，S/U/V (Silence/Unvoicd/Voiced)的判决和基音周期的准确检测还是一个公开的难题，其困难主要体现在：语音信号的时变性，背景噪声的影响，共振峰的影响，区别清音和低电平的浊音较困难，确定基音周期的起止点较困难。

基音周期检测方法大体上可以分为三大类：时域方法、频域方法和综合利用信号的频域和时域特性的方法。时域的方法直接处理语音信号的采样点，计算信号的波峰、波谷和过零率等。其特点是简单，计算量小，典型的方法是 Gold 和 Rabiner 提出的并行处理(PPROC)方法[9]。频域的方法主要计算信号的相关、功率谱和最大似然函数等，其精度要高于时域的方法，典型的方法有中央削波自相关法(AUTOC)[18]、平均幅度差分函数(AMDF)[39]法和倒谱法(CEP)[8]等。近些年来，又提出了一些精度更高、抗噪能力强的检测算法，但计算量都很大。考虑到我们处理的语音信号的信噪比较高，并且与文本无关的说话人识别主要基于语音的统计信息，有一些帧基音周期检测不准，对系统的影响不大，所以这里选用较简单的中央削波自相关法(AUTOC)[18]。

自相关法的原理是语音的短时自相关函数在基音周期的整数倍点上有很大的峰值。只要找到最大峰值点的位置，便能估计出基音周期。但实际上并不这么简单，最大峰值点的位置有时并不能同基音周期相吻合。产生这种情况的原因主要有两个，第一是窗的长度太短，二是因为声道的共振峰特性的干扰。为了克服这个困难，可以从两条途径入手。第一条是首先将语音信号通过 60Hz～900Hz 带通滤波器或 0Hz～900Hz 的低通滤波器。之所以高端截止频率定为 900Hz，是因为既可以去除大部分共振峰的影响，又可以当最大基音频率为 450Hz 时仍可以保留其一、二次的谐波。加低端截止频率可以抑制 50HZ 电源干扰。第二条途径，是将通过低通滤波器后的信号再进行非线性处理。中央削波即是一种有效的非线性处理方法。削波函数如图 2.2 所示。

$$y(n)=C(n)=\begin{cases}x(n)-L & \text{当 } x(n)>C_L\\ 0 & \text{当 } |x(n)|\le C_L\\ x(n)+L & \text{当 } x(n)<-C_L\end{cases}$$

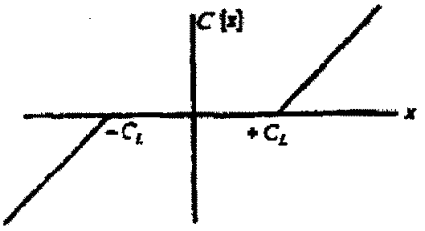


图 2.2 削波函数

削波电平 C_L 是这样确定的：将输入语音数据分为 3 个等长的子帧，分别在第一和第三子帧中寻找最大波峰值，削波电平取为两个峰值中较小的峰值和一个比例因子的乘积。

可以用一个能量门限来进行浊音的判决，也可以用估计出来的基音周期的连续性进行浊音的判决。

估计出来的基音周期轨迹可能同真实的基音周期轨迹在大部分段落是吻合的，而在一些局部段落中有一个或几个基音周期估值偏离了正常的轨迹(通常是偏离到了正常值的 2 倍或 0.5 倍)，为了去除这些点，可以采用各种平滑算法。

提取基音周期的步骤为：

(1)将输入语音信号通过低通滤波器;

(2)估计基音周期;

(3)平滑估计出来的基音周期。

2.2.2 线性预测系数

信号处理中系统传递函数参数模型主要有三种：第一种是只有零点没有极点的滑动平均模型(MA)，第三种是只有极点没有零点的自回归模型(AR)，第三种是既有零点又有极点的自回归滑动平均模型(ARMA)。考虑到声道的反射作用，其精确的模型应该是一个 ARMA，但 ARMA 的参数求解繁杂，在应用场合不宜实现。AR 模型能够很好地近似声道模型，且其参数求解相对容易，有多种解法，如针对自相关方程的 Durbin 递推算法和 Schur 递推算法、针对协方差方程的乔里斯基算法、Burg 算法，所以一般用一个 AR 模型来表示声道模型。

$$H(z)=G(z)V(z)R(z)=\frac{G}{A(z)}\frac{G}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{2.3}$$

其中 G 为增益系数，在下面的分析中不对其进行考虑。由上述传递函数可得到有关信号 $S(n)$的差分方程：$S(n)=\sum_{i=1}^{p}a_i S(n-i)+Gu(n)$，$S(n)$的线性预测$\vec{S}(n)$可近似表达为：$\vec{S}(n)=\sum_{i=1}^{p}a_i S(n-i)$，预测信号$\vec{S}(n)$的误差为：$e(n)=S(n)-\vec{S}(n)=Gu(n)$，线性预测$\vec{S}(n)$的传递函数为：$P(z)=\sum_{i=1}^{p}a_i z^{-i}$，误差 $e(n)$是信号 $S(n)$通过如下系统而产生的：$A(z)=1-\sum_{i=1}^{p}a_i z^{-i}$

由上我们可知$A(z)$约是声道模型传递函数的逆滤波器。$H(z)$参数可通过使$e(n)$在均方误差最小的准则下求得，求解算法有多种，具体解法可参见文献[1]-[7]。

线性预测分析从人的发声机理入手，通过对声道的短管级联模型的研究，认为系统的传递函数符合全极点数字滤波器的形式，从而 n 时刻的信号可以用前若干时刻信号的线性组合来估计。通过使实际语音的采样值和线性预测采样值之间达到均方差最小(LMS)，即可得到线性预测系数(LPC, Linear Predictor Coefficient)。对 LPC 的计算方法有自相关法(Durbin 法)、协方差法、格型法等等。计算上的快速有效保证了这一声学特征的广泛使用。

2.2.3 美尔倒谱系数

美尔倒谱系数也称感知频域倒谱系数(Mel-Frequency Cepstral Coefficients，简称 MFCC)，MFCC 分析着眼于人耳的听觉机理，依据听觉实验的结果来分析语音的频谱，获得了较高的识别率和较好的噪声鲁棒性。美尔倒谱系数利用了听觉系统的临界带效应，描述了人耳对频率感知的非线性特性[12,15]。

音高是一种主观心理量，是人类听觉系统对于声音频率高低的感觉。音高的单位是美尔(Mel)。响度级为 40Phon，频率为 1000Hz 的声音的音高定义为 1000Mel，16000Hz 的声音的音高为 3400Mel。

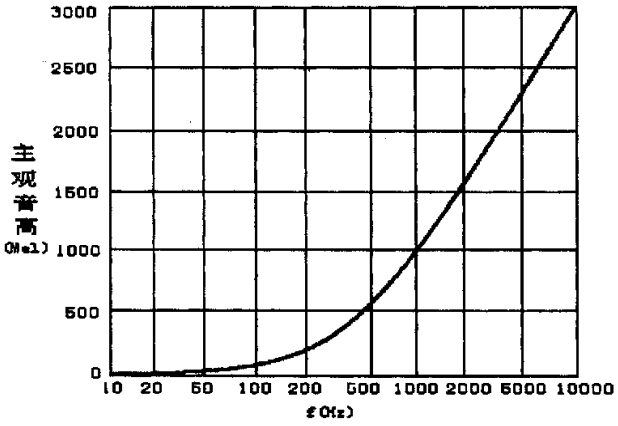


图 2.3 Mel 刻度与实际频率的关系曲线

图 2.3 就是主观音高与实际频率的关系曲线，它与 koening 频率刻度的趋势是很接近的。实际频率与 Mel 刻度频率之间的对应关系如下面公式所示：

$$mel(f) = 2595 * \log_{10}(1 + f/700) \qquad (2.4)$$

为了解释 MFCC 的提取过程，首先要解释一下临界频带(Critical-Band)的概念。研究发现，在声压恒定的情况下，当噪声被限制在某个带宽内时，其对人耳感觉的主观响度是恒定的，而一旦噪声突破了这个带宽，则主观响度的变化便会被感知。同样地，当声音恒定时，在这个带宽内的一个具有复杂包络的信号的响度等价于在这个带宽中心频率位置的一个纯音的响度，而与信号本身的频率分布无关：但是当信号的带宽突破了临界带宽时，其响度便不再等价。根据前人的工作，临界带宽随着频率的变化而变化，并与感知频率(Mel 频率)的增长一致。在 1000Hz 以下，大致呈线性分布，带宽为 100Hz 左右：在 1000Hz 以上带宽呈对数增长。根据临界带的划分，可将语音频域划分成一系列三角形的滤波器序列.即 Mel 滤波器组(如图 2.5 中所示)，取每个临界带内所有信号幅度加权和作为某个临界带滤波器的输出，然后对所有滤波器输出作对数运算，形成一个矢量，然后作离散余弦变换即得到 MFCC。

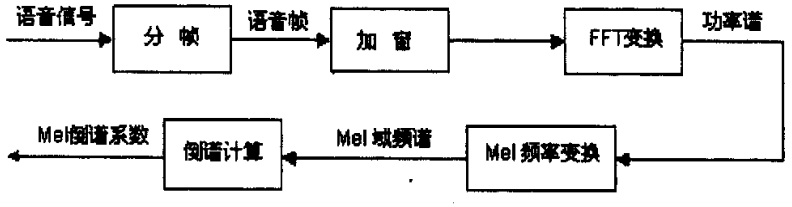


图 2.4 MFCC 的提取过程

一般取临界带滤波器组中滤波器的个数 D=20，所覆盖的最高频率 5.8KHz，MFCC 参数的提取过程可以参照图 2.4，其中 m 为帧标号，N 为一帧内的采样点数。MFCC 从人耳对频率高低的非线性心理感觉角度反映了语音短时幅度谱的特征，识别性能和抗噪性能均明显优于传统的线性预测倒谱参数 LPCC，是目前国内外非常流行使用的一种用于说话人识别的特征参数。

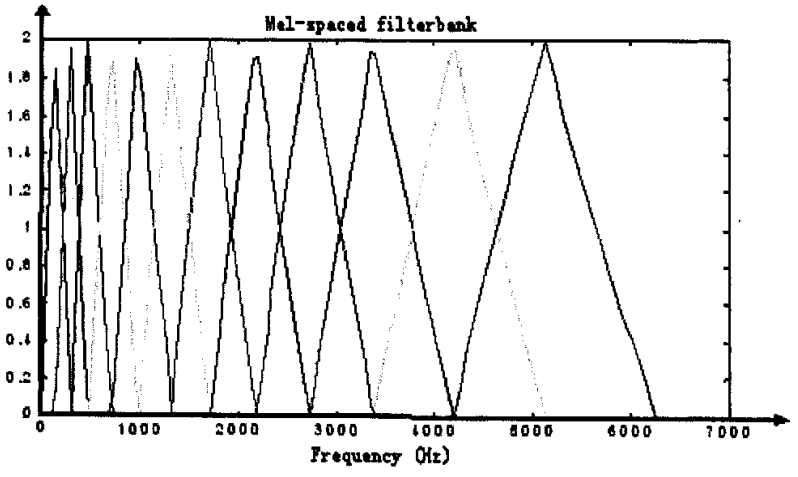


图 2.5 Mel 空间滤波器

2.3 说话人识别方法

2.3.1 识别方法分类

说话人识别技术可以分为三大类。最早的方法是声学特征的长时平均。其基本思想是频谱表示或基音等声学特征经过长时平均后便滤除了语音变化对声学特征的影响，剩下的当然就是与说话人相关的部分了。对于谱特征来讲，长时平均后便消除了语音变化对声学特征的影响，剩下的当然就是与说话人相关的部分了，一长时平均代表了说话人的声道状态。这种方法类似于高斯分类器，在一些比较困难的与文本无关的说话人辨认系统中应用得比较成功，然而这种平均方法丢失了太多的说话人相关信息，必须有较长(>20s)的语音才能获得稳定的长时语音统计。

第二类方法是为对应语音内容的说话人相关特征建模。识别时，将测试语

句中的语音声学特征和特定说话者的包含相同语音内容的模型相比较，这种比较主要体现的是说话人差异。一般的与文本有关的说话人识别方法比如 DTW、HMM 等都可归入这一类。而对于无限制文本来讲，必须在训练或识别之前进行语音切分，不管这种切分是显式的还是非显式的。显式的切分可以用一个基于 HMM 的连续语音识别系统作为前端处理，但这种切分几乎没有带来性能的提高，但却增加了计算的复杂度。非显式的切分方法则依赖于非监督的聚类，这种聚类不必给出每类的语音内容，因而对训练来讲不必切分。基于模板的匹配方法，比如矢量量化和 K 最邻近原则都可归入这一类。VQ 方法在限定说话者使用较小词汇(比如数字)时，效果相当好，但由于码书大小的限制，不易直接扩展到无限制文本的情况。和语音识别相似，概率模型法能较好地为声学特征建模并有一定的处理噪声和信道变异的能力，因此 HMM 及其各种变化形式在文本相关及无关的说话人识别中都获得很好的应用。与文本相关的说话人识别中，HMM 与语音识别中的应用基本一致。在与文本无关的说话人识别任务中，去掉 HMM 中的状态转移概率对识别没有影响，同时，在说话人识别中使用的 HMM 模型结构一般采用各态历经。

第三种方法也是最新的方法就是利用神经网络。神经网络不是为每个说话人训练一个模型，而是训练出一个判决函数来区分一个训练集内的不同说话人。多层感知器(MLP)、时延神经网络(TDNN)[12] ,径向基函数网络(RBF)及其改进方法[13]都在说话人识别中获得很好的应用。另外，文献[14]研究了使用多项式分类器的说话人识别，具有较高的识别性能。一般地讲，神经网络要比每个说话人有一个独立的模型需要的参数少，且识别性能也好，与 VQ 相当。其主要缺点是，对大多数神经网络来讲，当需要增加一个新的说话人时，整个网络要重新训练。

目前，VQ, DTW, GMM, HMM, ANN 等方法都被说话人识别广泛使用。文献[15]对 DTW, GMM, ANN 的说话人识别进行了性能比较。下面几节分别对几种主要的识别技术做比较分析。

2.3.2 基于 DTW 的说话人识别

动态时间规整(Dynamic Time Warping)是采用动态规划技术(Dynamic Programming)将一个复杂的全局最优化问题转化为很多的局部最优化问题进行分步的决策。

具体的思想是，设参考模式特征矢量序列为 $R=\{r_1,r_2,..r_I\}$,输入的待识别语音特征矢量序列为 $R=\{r_1,r_2,..r_J\}$，其中 $I \neq J$,DTW 算法就是要寻找到一个最佳的时间规整函数，使得待识别语音模式的时间轴 j 非线性的映射到参考模式的时间轴 i，最终的目标是使总的累计失真量最小，如图 3.1 为时间归整过程示意图，图中的格点为参考模式和测试模式的交会点，要找到一条若干交会点的路径使得总的失真最小。

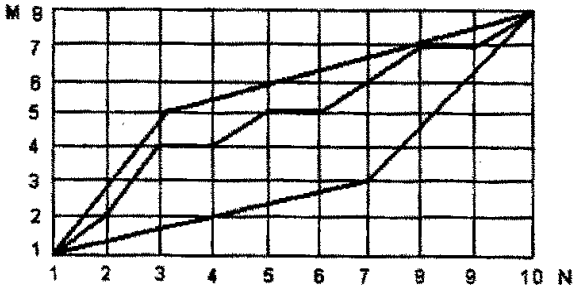


图 2.6 时间规整过程

设时间规整函数为： $C=\{c(1),c(2),...c(N)\}$，其中 N 为匹配路径长度，$c(n)=(i(n),j(n))$,表示第 i 个匹配点对是由参考模式的第 $i(n)$ 个特征矢量与待识别的模式第 $j(n)$ 个特征矢量构成的。两者间的失真值 $d(r_{i(n)},t_{j(n)})$ 称为局部匹配距离，DTW 算法就是通过局部优化的方法实现加权距离总和最小，即：

$$D=\min_{c}\left(\frac{\sum_{n=1}^{N}[d(r_{i(n)},t_{j(n)})*W_n]}{\sum_{n=1}^{N}W_n}\right) \qquad (2.6)$$

加权函数的选取考虑两个因素：1、根据第 n 对匹配点前一步局部路径的走向来选取，乘 45 度方向的局部路径，以适应 $I \neq J$ 的情况。2、考虑语音各部分给不同的权值以加强某些区别特征。为了保证匹配路径不违背语音信号各部分特征的时间顺序，对规整函数需要做以下约束：

(1) 单调性： $i(n)\geq i(n-1), j(n)\geq j(n-1)$

(2) 起点和终点的约束： 一般是要求 $i(1)=j(1)=1; i(N)=I; j(N)=J$

(3) 连续性：一般路径不允许跳过任何一点，即：

$$i(n)-i(n-1)\leq 1 \qquad j(n)-j(n-1)\leq 1$$

(4) 最大的规整量不超过某一极限，即 $|i(n)-j(n)|<M$， M 称为窗宽，另外还

会对搜索区域进行限制，如限制在平行四边形内等。

基于上述的这些概念，给出 DTW 算法的基本步骤：

首先，定义最小累计失真函数 $g(i,j)$，函数表示到匹配点对(i,j)为止前面所有可能的路径中最佳路径的累计匹配距离，有如下的等式，

$$g(i,j)=\min_{(i1,j1)\to(i,j)}\{g(i1,j1)+d(a_i,b_j)w_n\} \tag{2.7}$$

$(i1,j1)$表示局部路径$(i1,j1)\to(i,j)$起点，w_n为路径的权值，与局部路径的选取有关。

(1)初始化：令 $i(1)=j(1)=1$，$g(1,1)=2d(a_1,b_1)$

$$g(i,j)=\begin{cases}0 & \text{当 } (i,j)\in R\\ huge & \text{当 } (i,j)\notin R\end{cases} \tag{2.8}$$

R 为平行四边形的约束区域，顶点为$(1,1)$和(I,J)。

(2)递推求累计距离：

$$\begin{aligned}g(i,j)=\min\{\ & g(i-1,j)+d(a_{i-1};b_j)*w_n(1);\\ & g(i-1,j-1)+d(a_i;b_j)*w_n(2);\\ & g(i,j-1)+d(a_i;b_j)*w_n(3)\}\\ (i=2,3,\ldots\ldots\ & I;\ j=2,3,\ldots\ldots\ J;(i,j)\in R)\end{aligned} \tag{2.9}$$

最终的加权距离一般要用$\sum w_n$来补偿，当加权函数取得合适时，有

$$\sum w_n=I+J \tag{2.10}$$

因此最终距离为：

$$D=g(I,J)/(I+J) \tag{2.11}$$

(3)回溯求出匹配点对，根据上面求出的路径，由(I,J)向前回溯到起点$(1,1)$。不过该过程对于识别并没有必要，得出匹配距离即可，只有在求聚类中心时才必须回溯。

DTW 算法能够保证参考模式和待识别的模式沿着时间轴动态的匹配，实现最优非线性时间对齐，使得匹配的距离最小，距离最小的参考模式所对应的类就是识别的结果。

2.3.3 基于 VQ 的说话人识别

矢量量化的基本原理：将若干个标量数据组成一个矢量(或者是从一帧语音数据中提取的特征矢量)在多维空间给予整体量化，从而可以在信息量损失较小的情况下压缩数据量。

量化区间对应于胞腔(Voronoi cell)，胞腔是多维空间中的一个区域，量化值则对应于量化矢量，它是各个胞腔的形心。设矢量维数为 K，则 N 个胞腔各有一个 K 维的量化矢量，即 $y_1 y_2 \cdots y_b \cdots y_N \in R^k$。量化矢量也称为码字，这 N 个码字的集合则称为一个码本。显然，对于编码输出为 b 比特二进制数的矢量量化器，其码木大小为 $N=2^b$,即码本为：$Y_N=\{y_i, i=1,2,...,N\}$

利用矢量量化技术时，设计一个好的码本是很重要。这关键是如何划分 N 个区域边界。这需要用大量的输入信号矢量，经过统计实验才能确定。这个过程称为“训练”，它的任务是建立码本。它应用聚类算法，按照一定的失真准则，对训练数据进行分类，从而把训练数据在多维空间中划分成一个个以形心(码字)为中心的胞腔，常用 LBG 算法来实现。下面给出以欧氏距离计算两个矢量之间的畸变时，LBG 算法的框架[6]。

(1)将形成 VQ 码本所需全部输入矢量 X 存储于计算机内存中。全部 X 的集合用 S 表示。

(2)设置迭代算法的最大迭代次数 L。

(3)设置畸变改进阈值 δ。

(4)设置 M 个码字的初值 $Y_1^0, Y_2^0, ...Y_M^0$

(5)设置畸变初值 $D^{(0)}=\infty$

(6)设置迭代初值 $m=1$。

(7)根据最近邻准则将 S 分成 M 个子集 $S_1^{(m)}, S_2^{(m)}, ..., S_M^{(m)}$即当 $X \in S_1^{(m)}$ 时，下式应成立：　$d(X, U_1^{(m-1)}) \le d(X, Y_i^{(m-1)}), \forall i, i \ne l$

(8)计算总畸变 $D^{(m)}$：　$D^{(m)}=\sum_{i=1}^{M} \sum_{X \in S_i^{(m)}} d(X, Y_i^{(m-1)})$

(9)计算畸变改进量 $\Delta D^{(m)}$ 的相对值 $\delta^{(m)}$：　$\delta^{(m)}=\frac{\Delta D^{(m)}}{D^{(m)}}=\frac{|D^{(m-1)}-D^{(m)}|}{D^{(m)}}$

(10)计算新码字 $Y_1^{(m)},Y_2^{(m)},\ldots,Y_M^{(m)}$：　$Y_i^{(m)}=\frac{1}{N_i}\sum_{X\in S_i^{(m)}}X$

(11)$\delta^{(m)}<\delta$?

若回答为是，转入(13)执行

若回答为否，转入(12)执行

(12)m<L?

若回答为否，转入(13)执行。

若回答为是，令 $m=m+1$，转入(7)执行。

(13)迭代终止，输出 $Y_1^{(m)},Y_2^{(m)},\ldots,Y_M^{(m)}$作为码字，并且输出总畸变 $D^{(m)}$

(14)结束。

对于上列算法，需要做一些说明和进一步的讨论。第一，为了使迭代计算不致无限循环下去，设置了 δ 和 L 两个阈值参数。δ 的值远小于 1，当 $\delta^{(m)}<\delta$ 时，表明再进行迭代运算畸变的减小是极有限的，这时可停止运算。L 是限制最大迭代次数的参数，以防止 δ 设置的较低时迭代次数过多。第二，此算法的关键是第(7)和第(10)两项。第(7)项的工作是以第 $(m-1)$步形成的 M 个码字 $Y_i^{(m-1)}$ 为基准，将全部 X 的集合按照最近邻准则划分为 M 个子集 $S_i^{(m)},i=1\sim M$。每一个子集可以看成一个小区，在模式识别理论中称为“聚类区”。由此形成的划分一般也称为 Voronoi 划分，对于 $Y_i^{(m-1)}$而言，它所给出的总畸变 $D^{(m)}$是最小的。第(10)项完成的工作是按照第(7)项得到的 Voronoi 划分求出新的码字 $Y_i^{(m)}$。当采用欧氏距离来计算畸变时，$Y_i^{(m-1)}$应是 $S_i^{(m-1)}$中所有矢量的质心。由于 $Y_i^{(m-1)}$不一定是 $S_i^{(m)}$矢量的质心，用 $Y_i^{(m)}$替代 $Y_i^{(m-1)}$，必然能使总畸变下降。下一轮迭代计算中，以 $Y_i^{(m)}$为基准形成新的 Voronoi 划分 $S_i^{(m+1)}$时，总畸变显然又低于前一步的划分 $S_i^{(m)}$。这样，每完成一次迭代计算，总畸变必然有所降低。因此这个算法是一种使总畸变单调下降的算法，按照 Voronoi 划分，一个 VQ 系统的总畸变是它的 M 个码字决定的状态空间点的函数。如果这是一个凸函数，也就是说此函数只有一个全局最小点而没有局部最小点，那么这一使总畸变单调下降的算法将使迭代计算得到的解收敛到全局最小点上。然而在绝大部分实际情况中，该函数并非凸函数，即有全局最小点又有多个局部最小点。迭代算法的解收敛到哪个最小点取决于 M 个码字初值。虽然随即将给出若干初值设置的方法，但没有一种方法能够保证能收敛到全局最佳解。一种解决的方法是设置多组不同的初值，分别进行迭代计算，从中找出一个最佳解。这虽然增加了得到最佳解的机会，但是计算量庞大且不能保证必然获得最佳解。彻底解决的方法是采用模

拟退火的算法，其代价是付出非常大的计算量。

应用 VQ 的说话人识别有两个步骤：一是利用每个说话人的训练语音，建立参考模型码本，二是对待识别话者的语音的每一帧和码本码字之间进行匹配。由于 VQ 码本保存了说话人个性特征，这样我们就可以利用 VQ 法来进行说话人识别。在 VQ 法中模型匹配不依赖于参数的时间顺序，因而匹配过程中无需采用 DTW 技术，而且这种方法比用 DTW 方法的参考模型存储量小，即码本小。

我们可以将每个待识别的说话人看作是一个信源，用一个码本来表征，码本是从该说话人的训练序列中提取的特征矢量聚类而生成，只要训练的数据量足够，就可以认为这个码本有效的包含了说话人的个性特征，而与说话的内容无关。识别时，首先对待识别的语音段提取特征矢量序列，然后用系统已有的每个码本依此进行矢量量化，计算各自的平均量化矢量。选择平均量化矢量最小的那个码本所对应的说话人作为系统识别的结果。

应用 VQ 的说话人识别过程的步骤如下：

1.训练过程

(1)从训练语音提取特征矢量，得到特征矢量集;

(2)通过 LBG 算法生成码本;

(3)重复训练修正优化码本;

2.识别过程

(1)从测试语音提取特征矢量序列 X_1，X_2，…，X_n

(2)由每个模板依次对特征矢量序列进行矢量量化，计算各自的平均量化误差：$D_i = \frac{1}{M}\sum_{m=1}^{M}\min_{1\le l\le L}[d(X_n, Y_l^i)]$

式中，Y_l^i, $l=1,2,..,L$, $i=1,2,..N$ 是第 i 个码本中第 l 个码本矢量，而 $d(X_n Y_l^i)$是待测矢量 X_n 和码矢量 Y_l^i 之间的距离;

(3)选择平均量化误差最小的码本所对应的说话人作为系统的识别结果。

VQ 方法已成功应用于说话人识别中，是目前文本无关的说话人识别方法的评估基准。近期研究主要集中在 VQ 的改进算法中[16][17]。

2.3.4 基于隐马尔可夫模型的说话人识别

早在 1960~1970 年间，Baum 和他的同事就发表过多篇文章阐述了 HMM 的基本理论，但因为 HMM 理论大多发表在数学杂志上，且理论叙述不很详细不

便于理解，所以工程人员一般不太感兴趣读，因此 HMM 理论只初步应用到语音信号处理当中。80 年代后期，随着有关 HMM 理论详尽叙述的展开和一些指导性文章的发表，以及 HMM 模型参数最优化估计方法的解决，HMM 理论开始广泛的应用到语音信号处理（Speech Processing）当中。

HMM 应用概率统计的方法来描述时变语音信号，同时它可以很好的描述语音特征统计分布的统计模型，是准平稳时变语音信号分析和说话人识别有力的工具[1-8]。语音信号的不确定性，说明了它具有统计的确定性。为了描述这种语音信号随时间变化的特性，采用“状态”的概念，语音特征的变化表现为从一个状态到另一个状态的转移，即使是同一说话人的不同次发音，这种变化也只是统计确定的，表现为：特征从一个状态到另一个状态只是依一定的概率转移；处于某一状态时只是依一定的概率或概率密度获得语音特征。

描述该模型的参数主要有初始概率分布 a，状态转移概率矩阵 A，状态生成概率矩阵 B。 HMM 模型可分为离散 HMM，连续 HMM，半连续 HMM，和高斯混合 HMM 等几种。

应用 HMM 模型进行说话人识别时，针对每一个说话人的语音信号提取特征矢量，然后为每一个说话人建立一个 HMM 模型，$\lambda^i=(a^i,A^i,B^i)$为第 i 个说话人的模型参数。识别时计算未知语音信号的特征矢量 $\overline{O}$ 以及概率 $P(\overline{O}|\lambda^i),i=1,...,N$。对于说话人辨认，其中概率 P 最大的模型 λ 对应的说话人为识别结果；对于说话人确认，将计算得到的 P 值与已确定阈值相比较，小于阈值拒绝，大于阈值接受。

2.3.5 基于人工神经网络的说话人识别

说话人识别包含着从低层次到高层次的各个阶段及其彼此之间的相互作用，这是一个非常复杂的模式识别过程，而模式识别的最新技术——人工神经网络，尤其适合于此类问题。其中较为成功的例子多数集中在说话人个性特征抽取这一层次上，用于说话人识别的神经网络结构集中在多层感知器结构的神经网络，如反向传播人工神经网络（BP-Back Propagation Network）、人工神经预测网络(NP-Neural Prediction)、径向基函数神经网络(RBF-Radial Basis Function)、时间延迟人工神经网络(TDNN-Time Delay Neural Network)等。与传统的说话人识别方法相比，人工神经网络的出现和发展为说话人识别开拓了新的思路，它通过

人工神经网络强有力的自适应、自学习和自组织能力实现对说话人语音信号特征的分类和识别，其网络权值形成了说话人个性特征的隐式表示，是一种很有前途的识别方法。但网络训练速度、网络训练的收敛性以及识别系统的通用性等方面仍存在许多问题，沿着这一思路进行说话人识别的研究将依赖于人工神经网络理论的不断成熟和发展。

但是从总体上讲，基于人工神经网络法的说话人识别技术的研究目前还处于研究与实验阶段，对于应用方法的研究也刚刚起步，在说话人识别系统的应用方面也尚在摸索。

2.4 说话人确认和说话人辨认

说话人识别可以分为说话人确认和说话人辨认。说话人辨认用来确定待识别的语音是哪一位注册过的说话人说的，而说话人确认用来确定待识别的语音是否是说话人所宣称的那个人说的。如果在应用系统中，语音用来确认说话人所宣称的身份，那么应用系统归为说话人确认系统。说话人确认和说话人辨认的根本区别在于系统决策时可能结果的数目，说话人辨认中待定的数目等同于注册说话人集合的大小，而说话人确认只有两个可能的结果，接受或是拒绝，与说话人集合的大小无关。因此，说话人辨认系统随着说话人集合的增大性能下降，而说话人确认系统则能保持在一个恒定的水平。

如果对于一个待识别的说话人，集合中可能没有所对应的说话人模型，那么该系统为开集的说话人辨认系统，这种情况下，系统决策时需要增加一个可能的结果，即待识别的说话人不属于说话人集合。其实，无论是说话人确认还是说话人辨认，都可以用一个阈值检验来确定当前匹配是否有足够可信度，是接受还是要求再来一次。

根据识别方式，说话人识别又可以分为文本有关和文本无关两种方式。前者要求在识别时，说话人提供与训练语音文本相同的关键词串或者语句的语音，而后者则没有这样的要求。文本有关通常基于模板匹配技术，将待识别语音样本与说话人参考模板在时间上对齐，然后从头至尾累积计算样本和模板的相似度。因为直接利用了与每个音素或音节相联系的个体特征，所以通常系统识别性能要比文本无关的方式好得多。

第三章 说话人识别系统设计

3.1 识别算法的选择

在第二章中对各类说话人识别方法的阐述的基础上，下面在复杂度和识别率方面对各个方法作简单的比较[22]，表 3.1 中，CHMM 为连续隐马尔可夫模型，DHMM 为离散隐马尔可夫模型，都是基于 HMM 的改进方法。

DTW 模板匹配技术的缺点是只对特定语音识别有较好的识别性能，并且在使用前需要对所有词条进行训练。这一应用从 20 世纪 90 年代就进入成熟期。目前的努力方向是进一步降低成本、提高稳健性和抗噪性能。

基于 HMM 技术的识别系统可用于非特定人，不需要用户事先训练。它的缺点在于统计模型的建立需要依赖一个较大的语音库。这在实际工作中占有很大的工作量。且模型所需要的存储量和匹配计算的运算量相对较大，通常需要具有一定容量 SRAM 的 DSP 才能完成[29][30]。

人工神经网络 ANN 在语音识别领域的应用是在 20 世纪 80 年代中后期发展起来的。其思想是用大量简单的处理单元并行连接构成一种信息处理系统。这种系统可以进行自我更新，且有高度的并行处理及容错能力，因而在认知任务中非常吸引人。但是 ANN 相对于模式匹配而言，在反映语音的动态特性上存在缺陷。单独使用 ANN 的系统识别性能不高，所以目前 ANN 通常在多阶段识别中与 HMM 及 MFCC 算法配合使用[28]。

VQ 方法在较小数据量的情况下，依然能有较好的识别率[20,21]，而且实现的复杂性较低，适合于在嵌入式系统中的应用。

表 3.1 不同算法的复杂度及性能比较

	VQ	DTW	CHMM	VQ+DHMM	ANN
实现复杂度	简单	简单	复杂	复杂	中等
识别率	中等/高	低/中等	中等/高	中等/高	中等

3.2 确认阈值的选择

说话人确认中，是将待识别语音与注册说话人自己的模型比较，以确定是否是注册者本人的声音，系统只需给出接受或拒绝两种选择，因此，说话人确认系统的性能与说话人集合的规模无关。

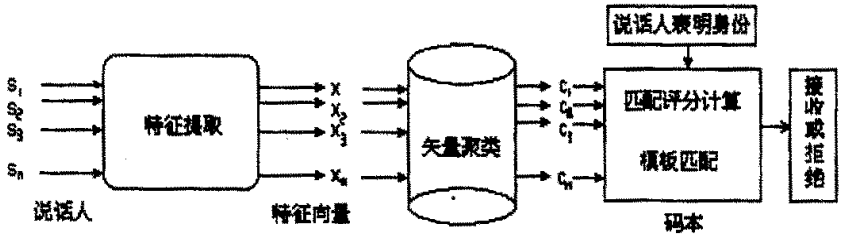


图 3.1 说话人确认系统框图

说话人确认系统中关键有三个步骤：

1、特征提取：这是说话人确认系统中非常关键的一个步骤。

不同说话人在特征参数的分布情况，即特征的相关性对系统性能有很大影响。如果两个说话人在特征空间的分布有很大的重叠，也就是说两个人很相似，系统的识别效果肯定不好，即两个说话人模板之间距离很近，这样就给确认造成很大困难，错误率将会提高。

2、说话人训练模板生成

从多个训练数据中提取特征，进行聚类分析，形成最能代表该说话人特征的训练模板，在测试过程中，就以此模板为标准，以输入数据的特征与训练模板之间的距离大小来做出接收或拒绝的确认。

3、说话人确认阈值

确认阈值的设计是说话人确认系统的关键问题之一，也是影响说话人确认系统实用化的难点所在。说话人确认实质上是一个二元判决问题，即说话人的确认语句与其说话人的参考模型的距离小于确认阈值时，系统予以确认；反之，则系统拒绝承认。确认阈值的设计是在训练说话人参考模型时完成的。由于事先不可能收集到所有冒认者的语音数据，所以确认阈值只能根据本人的数据和有限的冒认者数据来确定。这样，由于冒认者事先是未知的，实际使用时系统的性能总是下降，甚至下降很多。

确认阈值的设计是说话人确认系统的关键问题之一，是影响说话人确认系

统实用化的难点所在。因为通常情况下，各个不同说话人的语音特征参数空间均有部分是重叠的，事实上往往是一个多模式参数空间分割问题。在说话人确认系统中，要将待识别说话人的输入语音计算出的参数与其所声称的说话人的参考参数比较，如果二者的距离小于规定的阈值，则认为是声称说话人，否则认为不是声称说话人。问题是在于怎样确定阈值使系统的效果最佳以及对不同的说话人阈值如何调整。

由于缺少合适的数据，合理估计能准确反映说话人模型间变化的阈值就变得极为困难。阈值需要根据用户的训练数据和冒认者的数据来设置。通常阈值的设置都是利用统计特性的方法，我们认为说话人语音本人对于模型测度和冒认者对于模型测度都是随机变量，假定它们服从正态分布，其均值和方差分别是σ, m和σ', m'，如图 3.2。有以下几种方法。

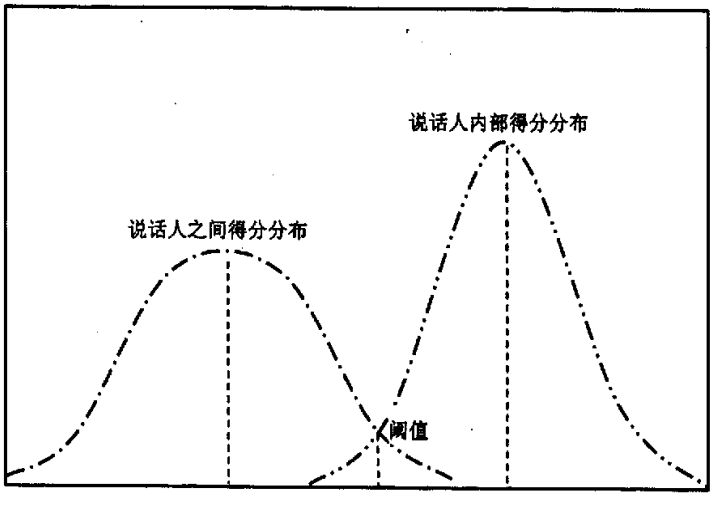


图 3.2 语音的模型测度分布图

1. 只利用均值

$\theta = \beta m + (1-\beta)m'$，其中，$\theta$是阈值，$\beta$为优化因子

2. 利用均值和方差，根据正态分布的 3σ 原则，99. 7%的样本应落在$(m-3\sigma, m+3\sigma)$区间内，再考虑到冒认的拒绝问题，可以如下设置阈值θ：

$$\theta = \begin{cases} m-3\sigma & m-3\sigma > m'+3\sigma \\ \dfrac{m\sigma'+m'\sigma}{\sigma+\sigma'} & else \end{cases} \tag{3.1}$$

根据需要，可以把3σ改为2σ或σ

3.利用似然比

对说话人判别的问题实际上是个假设检验的问题。设 $\lambda_0=\{m,\sigma\}$ 代表了说话人自身分布的参数，$\lambda_1=\{m',\sigma'\}$ 代表了冒认者的分布，则假设检验问题可描述为：

$$\begin{cases}H_0:\lambda=\lambda_0\\H_1:\lambda=\lambda_1\end{cases}$$

从而判决规则变为：拒绝 H_0，当且仅当 $\frac{P(O|\lambda_0)}{P(O|\lambda_1)}<\gamma$ 其中，γ 为优化因子。

3.3 开集说话人识别

一、闭集

假设说活人识别系统中已经训练的有 N 人。闭集的说话人识别定义为系统具有这样的先验知识：每一个测试者都是这 N 人之一。

闭集的说话人辨认测试需要对测试音对每一个说话人的模型进行计算，得到每一个人的相似度，系统认为相似度最大的说话人为识别结果。

二、开集

开集的说话人识别系统定义为测试者可能来自这 N 个人以外。与闭集测试不同，开集测试的时候，不管是说话人辨认还是说话人确认，都需要把模型计算出来的值与系统设定的阈值比较，如果大于系统的阈值，则判断为测试者为 N 人之外。这与说话人确认类似，在说活人确认系统中，也需要有一个阈值来决定接受还是拒绝该说话人。

阈值的设定对系统系能的影响主要体现在错误拒绝率和错误接受率的变化。一般来说，错误拒绝率随着系统阈值的提高而降低，错误接受率随着系统阈值的提高而升高。具体的阈值怎么设，可以根据实际需要。在某些宁可错误拒绝多次也不可错误接受一次的情形，阈值可以设得高一点，反之就可以设的低一些。

对于一个开集的说话人识别系统而言，可以将所有的参考说话人看作是一个集合，那么在进行识别的时候只需要判断待识别说话人是否属于该集合即可。如果待识别说话人是某个参考说话人，有时还需要给出具体的识别结果。根据集合的不同描述方法，开集的说话人识别有两种方式。当使用描述法表示集合

时，需要找到集合中元素的性质，这样在进行开集的说话人识别时，只需判断待识别说话人是否满足集合的性质即可。虽然这种识别方式的过程很简单，但是在通常情况下，很难找到一种准确的数学形式来描述所有参考说话人组成集合的性质，同时，这种方式只能判断说话人是否属于集内，而不能给出具体的识别结果。使用列举法表示集合，不仅可以判断未知语音是来自集内还是集外说话人，同时还可以得到具体的识别结果，因此，一般使

用列举法来表示由所有参考说话人组成的集合。利用列举法表示集合有两种识别方式：

1. 逐一识别

将待识别语音分别和每个说话人的模型进行相似度匹配，逐一判断待识别语音是否由集合中参考说话人发出。如果在判断的过程中，找到与待识别语音相匹配的参考说话人，则将该说话人作为最后的识别结果；如果在逐一判断的过程中没有找到相应的参考说话人，则认为待识别语音是由集外的说话人发出。在将待识别语音与每个参考说话人进行逐一匹配时，逐一判断的过程其实就是确认的过程。

假设集合内有 N 位参考说话人，如果待识别说话人不是集合内的参考说话人，那么系统要进行 N 次确认，这样当集合内的参考说话人数 N 很大时，要做出集外说话人的判决所需要的计算量和时间都会很大，因此，目前采用第二种方式进行开集的说话人识别。

2. 先辨认后确认

在开集说话人识别中，可能存在不属于任何参考说话人的未知语音，这样可能的判决数目就变成了 N+1 个，其中需要一个格外的判定，即待识别的说话人与集合内的任意模型都不匹配，因此开集的说话人识别可以看作是一个辨认和确认联合的过程[19] 。具体的过程可以见图 3.3。

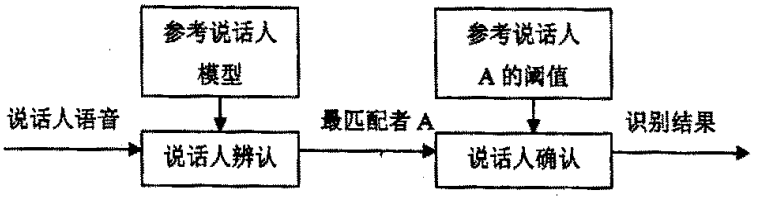


图 3.3 开集说话人识别框图

对于开集的说话人识别系统而言，在识别阶段，首先利用待识别语音和集内说话人的参考模型进行辨认，这是一个 N 选一的过程，找到与待识别语音最匹

配的参考说话人 A，作为辨认的结果，然后利用训练时产生的参考说话人 A 的阈值，对待识别语音进行说话人确认，得到最后的识别结果。

3.4 系统流程设计

本项研究将说话人识别系统从功能上分为七部分，就每一部分涉及的原理进行了说明及实现算法进行了分析。七部分包括：数据采集、预处理、划分语音段、语音特征提取、建模、模式匹配及推断模块(见图 3.4)

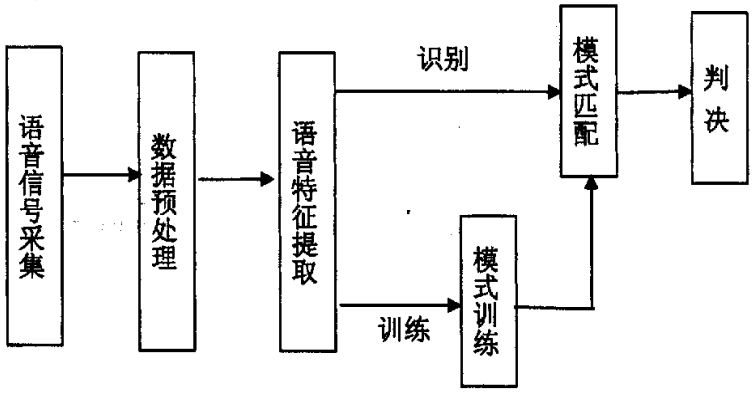


图 3.4 说话人识别系统的结构图

3.4.1 语音信号预处理

在对语音信号进行语音特征提取等处理以前，必须将之“标准化”，即预处理。预处理过程包括降噪、分帧及预加重等操作。

语音经声音采集设备，如麦克风，进行声电转换变为模拟信号，然后经由 A/D 进行采样、量化变为数字信号。对得到的数字信号进行信号能量的归一化处理以提高分析的稳定性，然后进行预加重以提升高频部分。此外由于对语音信号常采用短时分析技术，在语音信号分析之前，首先要对其进行分帧加窗，常用的窗函数有：矩形窗、汉明窗、汉宁窗等。

分帧时，每帧长度为 20ms 左右，帧与帧之间的偏移通常取帧长的 1/2 或 1/3，即每隔帧长的 1/2 或 1/3 进行分帧。分帧后是加窗，频域分析时常采用的是汉明窗，以减轻短时语音段边缘的影响。在分帧加窗的基础上即可对语音信

号进行语音分割和特征提取等处理。

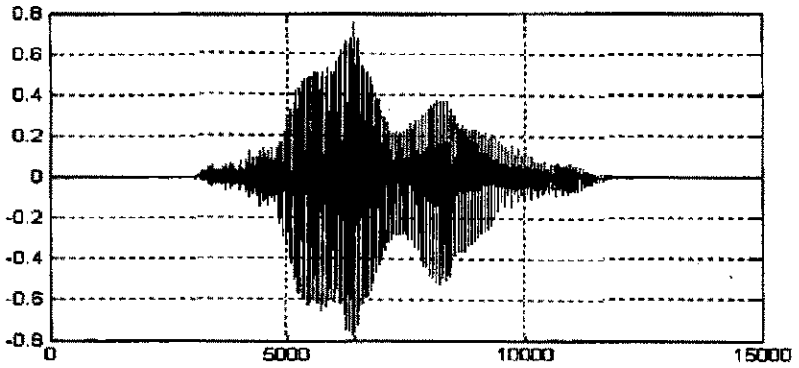


图 3.5 典型的语音信号时域图

一、语音信号获取

在 DSP 系统中，语音数据是实时采集实时处理。在仿真阶段，采用波形语音格式文件作为语音的输入，语音预先由软件录制，录音参数设置成单声道、采样精度 16bit 及采样频率为 8000Hz，录制长度为 2 秒。

二、语音端点检测

语音检测在语音处理中是一个很重要的方面。噪声环境中检测语音起止位置有利于提高语音系统性能，该节主要讨论了语音检测的一些常用方法，并着重提出了自相关法语音检测。

（1）语音和噪声的特性分析

噪声是一个随机过程，它甚至可以是时变的、非平稳的。非平稳噪声的参数将时时发生变化，使我们对其特性的估计变得困难，处理则更加的困难。这里我们假定噪声是平稳的，并且主要讨论加性噪声。

语音信号在时域上具有很强的时变特性。在有些段落中它具有很强的周期性、有些段落又具有噪声特性，而且周期性语音和噪声语音的特征也在不断变化之中，只有在较短的时间间隔中(如 20-200ms)才可以认为语音信号的特征基本保持不变。这一特点是语音信号数字处理的一个重要出发点。

（2）短时能量和短时过零率

由于语音信号幅度随时间有相当的变化，特别是清音段的幅度一般比浊音段的幅度小很多。所以用短时能量能够比较合适的反映这些幅度变化。通常，

定义短时能量为：

$$E = \sum_{m=-\infty}^{\infty}[X(m)W(n-m)]^2 \qquad (3.2)$$

语音信号是宽带信号，虽然平均过零率表示方法不那么确切，但是它还是能对语音频谱特性作粗略估计。短时过零率的计算公式为：

$$Z_n = \sum_{m=-\infty}^{\infty}|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]|\,w(n-m) \qquad (3.3)$$

$$w(n) = \begin{cases} 1/2N, 0 \le n \le N-1 \\ 0 \quad ,\text{其它} \end{cases}$$

实际应用中常常用到短时过零率的修正参数，它是一帧语音波形穿越某个非零电平的次数。此电平适当地设置为一个接近零的值时，对于清音仍然有很高的值，而对于无声则很低。

短时能量和短时过零率都是随机参数。不同性质的语音各自有不同的概率分布。对于静音、清音和浊音三种情况，浊音的短时平均幅度大而短时过零率最低；清音的短时平均幅度居中而短时过零率最高；静音的短时平均幅度最低而短时过零率居中。它们的条件概率分布都很接近于正态分布。

（3）利用短时能量和短时过零率检测语音信号端点

除非在高信噪比的声学环境中(如消声室或隔音室)的语音外，从背景噪声中鉴别语音不是一件简单的事。对于高信噪比环境，最低电平语声的能量(如弱摩擦音)超过背景噪声能量，简单的能量判断就可以得到比较满意得结果。

能量和过零率的方法是最简单的时域测量方法。首先，可根据浊音情况下短时能量的大小确定一个阈值 ITU。如果 ITU 的值定得比较高，一帧输入信号的短时能量超过 ITU 时，就可以十分肯定该帧信号不是无声，而有相当大的可能是浊音。根据 ITU 可判定输入语音的前后两个点 N_1 和 N_2, N_1 和 N_2 之间是语音段。但语音段的精确位置还要在 N_1 之前和 N_2 之后仔细查找。为此，再设一个更低的阈值参数 ITU,由 N_1 向前查找，当短时能量低于 ITU 时就可以确定点 N_1。类似地，由 N_2 向后找，可以确定 N_2。然后由 N_1 向前和 N_2 向后继续用短时过零率搜索。为此，根据无声情况下短时过零率均值设置一个阈值，根据上面方法找出语音的精确起止点。

为了避免将同一个语音段切分成两个或多个语音段，还需要在上面的检测基础上，对检出的语音段进行合并，当语音段的结尾处距离下一段的开始位置小于某个距离值时，可以将这两段语音合并成一段。另一方面，在语音边界检测

的过程中有时会遇到短时突发性的干扰噪声，它们的能量比较大，这样就会造成误判。这种干扰噪声持续时间很短，一般小于 50ms。为了消除这种干扰，本文用检测后的起止长度来判断它是不是语音。

下图显示了不同方法下端点检测的结果，在 DSP 实现中采用过零率来判断语音的起止点。

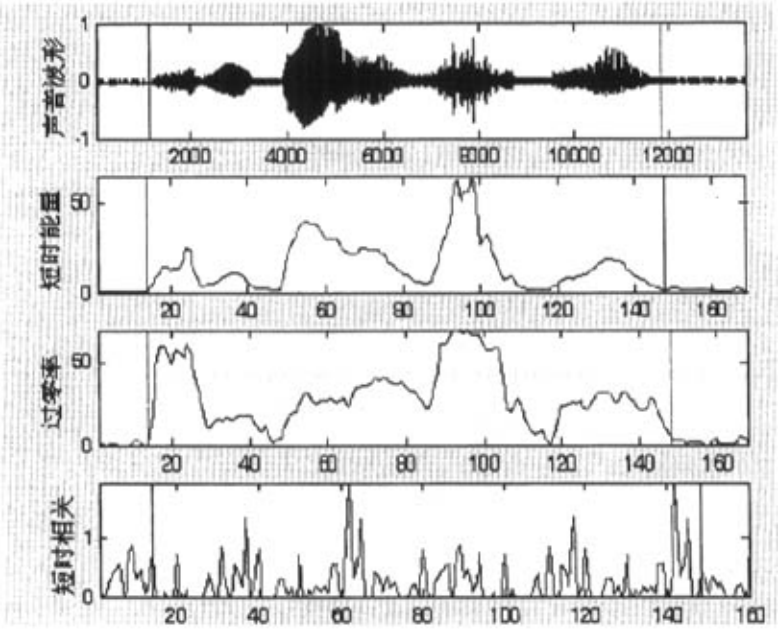


图 3.6 端点检测结果

三、信号预处理

预处理过程一般包括除噪声、分帧、预加重和加窗处理几个部分，图 7 描述了预处理过程。

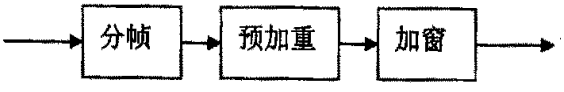


图 3.7 语音信号预处理过程

在语音信号处理中，用循环队列的结构来存储采样得到的数字信号，以便用一个有限容量的数据区来处理数据量很大的语音数据。在进行处理时，按一

帧接接一帧的方式从数据区取出，也就是将采样信号分为长度为 N 个样本的帧，并且相邻帧之间的位移为 M 个样本，参数 M 决定单位时间内的帧数。一般地，帧长取 25ms,帧移为 12ms。

对于语音信号的频谱，通常是频率越高衰减得越严重，因此必须对高频部分进行加重处理。经过预加重处理后的语音信号，其高频部分可与中频部分(1-2kHZ)的幅度相当，这个过程可以用公式表示：

$$S_{pe}(n) = S_{of}(n) - 0.97S_{of}(n-1) \qquad (3.4)$$

其中 S_{pe} 和 S_{of} 分别为预加重处理前后的信号。

最后，用汉明窗对原语音信号进行加窗处理。

3.4.2 特征参数的提取

特征参数的选取对于说话人识别的性能至关重要。在说话人识别的研究中，人们通过对语音信号时域和频域的分析，已提取出各种各样的语音参数。在算法的 DSP 实现过程中，选用了 Mel 倒谱系数作为说话人的特征参数。

Mel 倒谱系数的具体计算过程在第二章已经详细的讲解了，这里介绍一下使用 DSP 实现 Mel 倒谱系数的具体过程。在进行 Mel 倒谱系数的计算中，主要包括语音信号的预处理、短时功率谱的计算、功率谱的 Mel 域变换以及对 Mel 对数谱的 DCT 变换。其框图如图 3.8 所示。

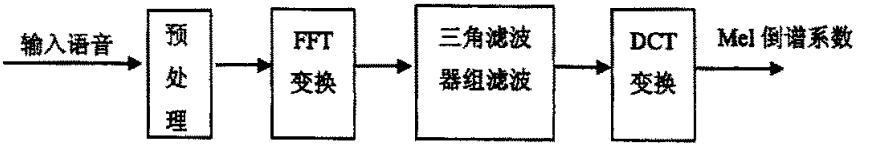


图 3.8 Mel 倒谱系数计算过程

1. 对连续的语音信号进行预处理，预处理的方法可参照第三章介绍的方法，得到经过预处理以后的若干帧短时的语音信号。

2. 计算每帧短时语音信号的短时功率谱。在计算功率谱的过程中，需要对每帧的短时信号进行 FFT 变换，在 DSP 中直接可以调用 DSP 库中自带的 FFT 库函数进行运算。然后计算 FFT 得到的复数序列模的平方，得到每帧短时语音的短时功率谱。

3. 利用三角滤波器组对每帧短时信号进行滤波，将语音信号变换为 Mel 频

率刻度。由于三角滤波器的中心频率是按 Mel 频率刻度均匀分布排列的，每两个相邻的滤波器的过渡带相互搭接，而且频率响应之和为 1，所以得到的滤波器组的系数矩阵 M 是一个稀疏矩阵（矩阵中每一行表示每个滤波器的系数）。通过对 M 的分析可知，对于每个滤波器而言，仅有连续的几个非零值，因此，为了节省 DSP 的存储空间，只存储每个滤波器中的非零值和非零值的坐标位置，来实现对语音信号短时谱的滤波.

4. 将 Mel 域上的功率谱取对数，得到 Mel 对数谱，然后再利用离散余弦变换（DCT）将其变换到时域，得到所要求的 Mel 倒谱系数。DCT 变换是离散傅立叶变换 DFT 的一种特例，是一种实偶的离散傅立叶变换。对于 DFT 变换而言，当序列的长度 N=2^n 时，可以使用 FFT 实现。所以利用 DSP 进行 DCT 变换，当 DCT 变换序列的长度为 N=16 时，直接使用 FFT 来实现 DCT 变换的计算。具体的实现步骤如下：

a.对处理的数据进行变形（设 $x(n)$为 N 点）

$$y(n)=x(2*n+1) \qquad n=0,1,...,N/2-1$$

$$y(n)=x(2*(N-n)) \qquad n=N/2,...,N-1$$

b.对变形得到的序列 $y(n)$进行 N 点的 FFT 得到 $yy(k)$;

c.令 $ww(k)=e^{-jkn/2N}*a_N(k)$ 其中 $a_N(k)=\begin{cases}\sqrt{\frac{1}{N}} & k=0\\ \sqrt{\frac{2}{N}} & k=1,...,N-1\end{cases}$

d.计算 $b(k)=yy(k)*ww(k)$;

e.对 $b(k)$取实部，得到所要计算的 Mel 倒谱系数。

下图为在 Matlab 仿真时得到的某两个说话人的特征矢量。不同的说话人的特征矢量有着明显的区别，这反映在曲线图上就是每个矢量有着不同的形态，经过矢量聚类过程对特征矢量进行压缩就得到说话人的码本。具体的提取过程将在下面给出。

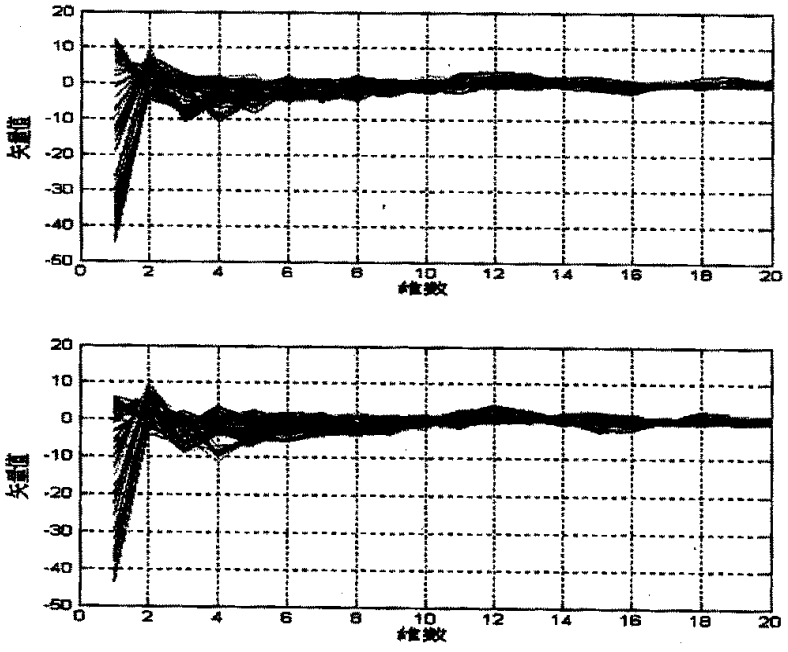


图 3.9 两个说话人语音的特征矢量

3.4.3 特征矢量聚类以及码本形成

特征矢量的聚类以及码本设计采用 LBG 算法，按最近邻准则用初始码本中的各个码字对训练序列进行 Voronoi 划分，从而形成 M 个子集，每一子集为一类，M 为码本容量；计算各类的形心和平均失真，迭代计算下去，不断对码本进行修正直到性能满足要求或不能再有明显改进为止。在 LBG 算法的 DSP 实现过程中，主要包括 3 个子程序的设计：初始码本子程序、失真距离子程序以及新码字求解子程序的设计。

1.初始码本子程序

在进行初始码本计算时，采用的是最大距离法。一段语音信号经过特征提取得到特征矢量序列 $\overrightarrow{X}=\{\overrightarrow{x_1},\overrightarrow{x_2},...,\overrightarrow{x_n}\}$；在特征矢量序列中随机选择一个矢量作为第一个码字 $\overrightarrow{y_1}$；将 $\overrightarrow{X}$ 中其余矢量与第一个码字相比较，找出失真最大的特征

矢量，作为第二个码字 $\overrightarrow{y_2}$ 。除前两个码字， $\vec{X}$ 中其余矢量与第二个码字相比较找出失真最大的特征矢量为第三个码字$\overrightarrow{y_3}$。以此类推，得到初始码本$\vec{Y}$。具体的流程见图 3.10 所示。

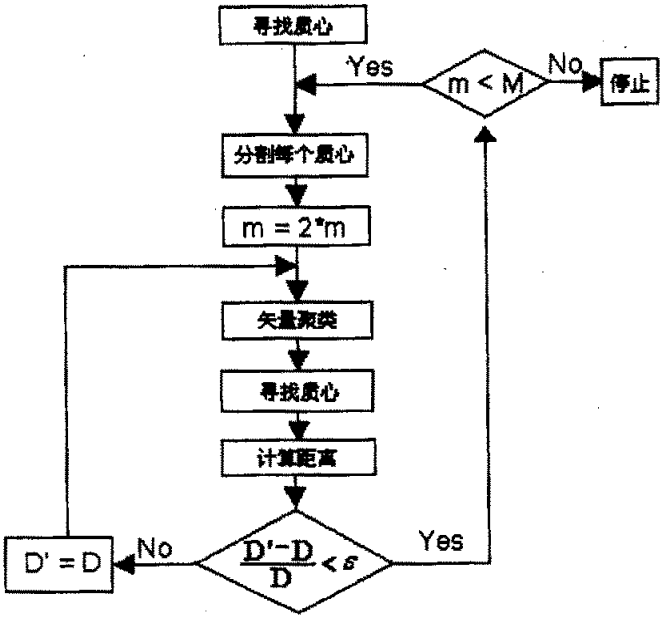
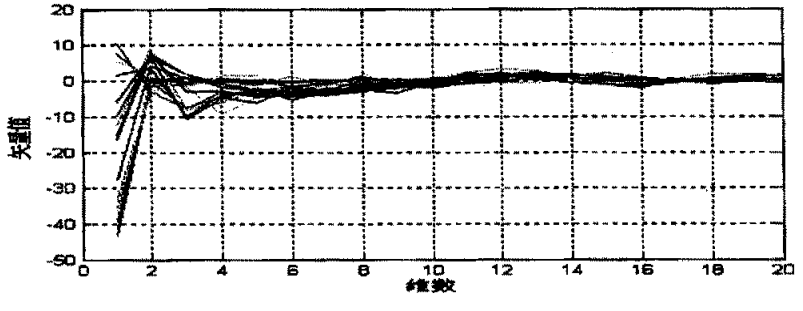


图 3.10 LBG 算法流程图

2. 测试者码本的生成

在仿真阶段，由 PC 机上预先生成测试者的码本，且码本作为常量固化到程序代码中。图 3.11 表征了两个说话人的特征码本。



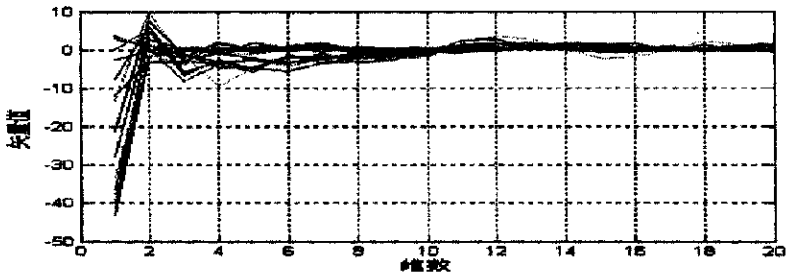
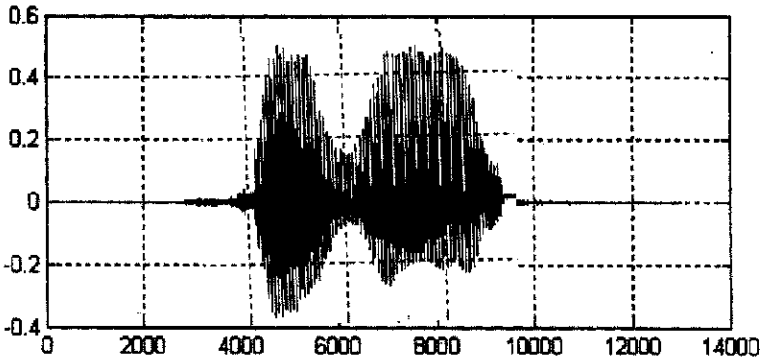


图 3.11 某两个测试者的码本

3.5 系统在不同参数下的性能分析

3.5.1 噪声环境测试



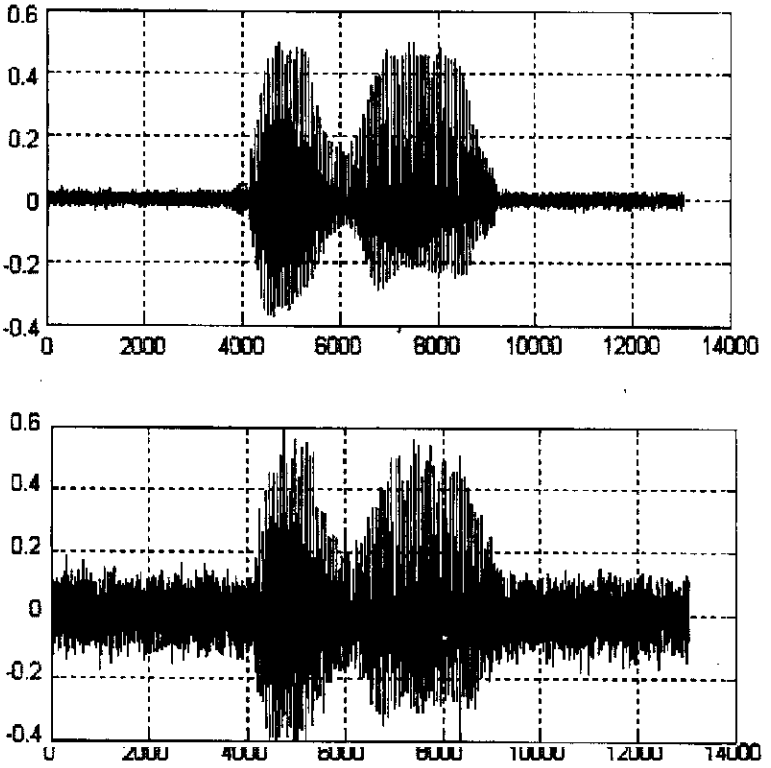
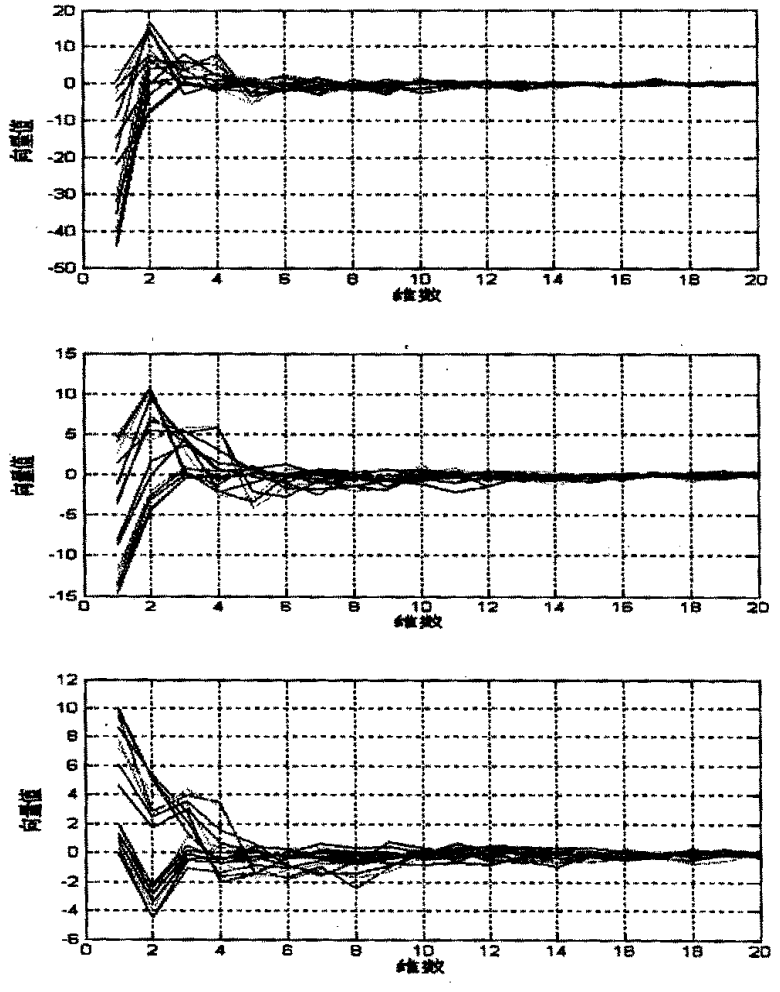


图3.12 原始语音与加噪后的语音

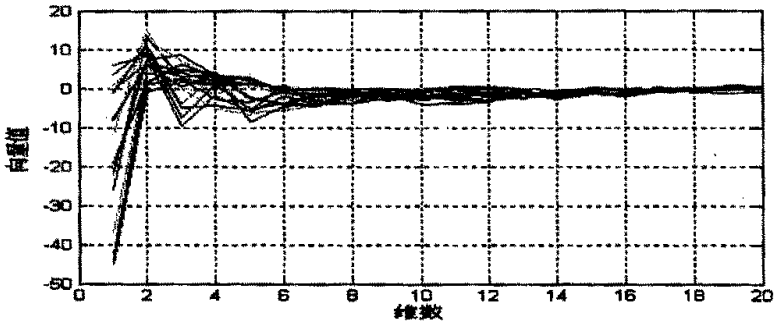
为了比较系统在不同噪声环境下的识别能力，采用往原始信号中加入白噪声的方法模拟噪声环境，加入的噪声方差有两个不同的值，分别为 0.0001 和 0.0025，以便比较。原始语音与加噪后的语音如图 3.12 所示，其中第一个图为原始信号，第二个为加入方差为 0.0001 的噪声后的语音信号，第三个的噪声方差为 0.0025。

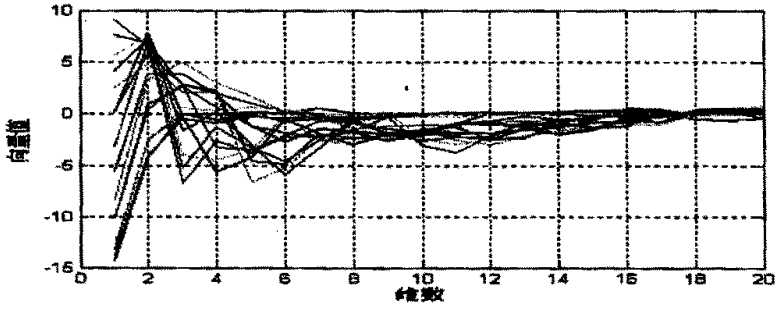
图 3.13 为三个语音样本加不同噪声前后计算出的码本对比，共三组，相邻三个为一组，横坐标为向量的维数，纵坐标为向量的值。计算出的码字是说话人的个体特征的反映，其中使用的语音的采样率为 12KHz。虽然码字的值前后有较大变化，但是依然能反映个体的语音特征，在仿真运行中，识别率有不同程度的降低(见表 3.2)。在实际运行中，系统在语音采集阶段必须对周围环境进

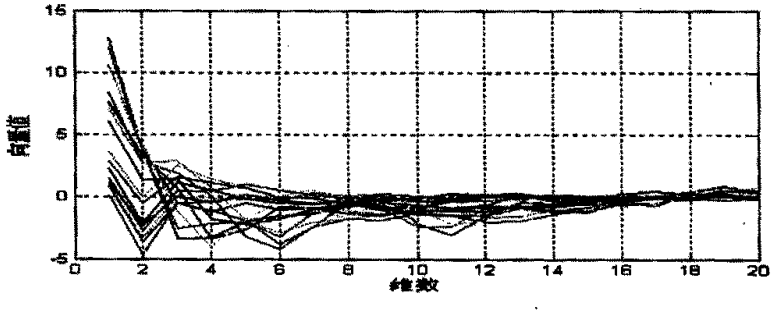
行适当控制，再辅以程序上的去噪，才能保证系统能有较好的识别率。



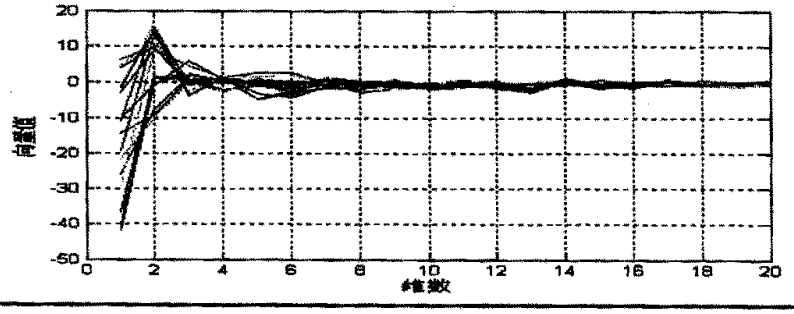
（组 1）

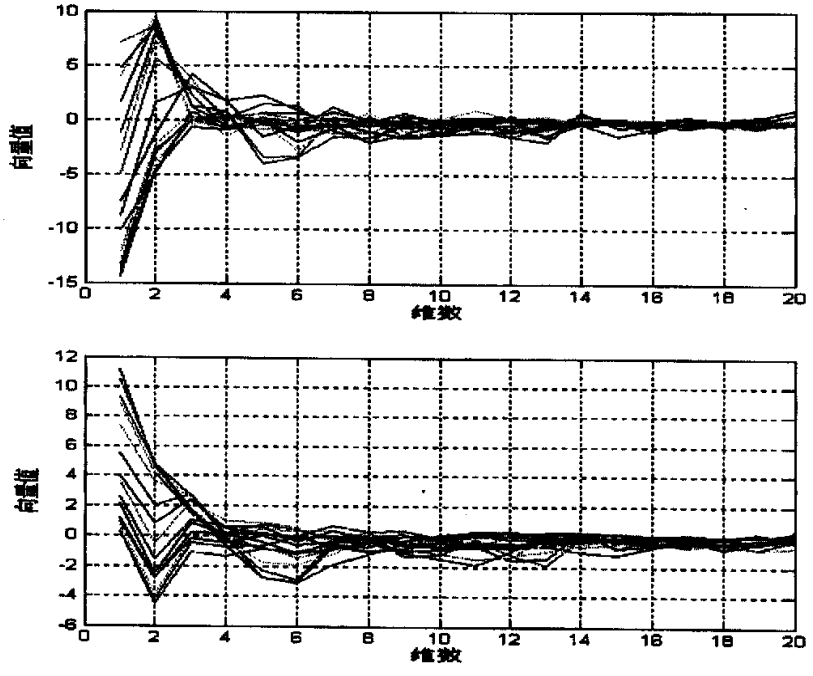






（组 2）





（组 3）

图 3.13 三个语音样本加噪前后的码本对比

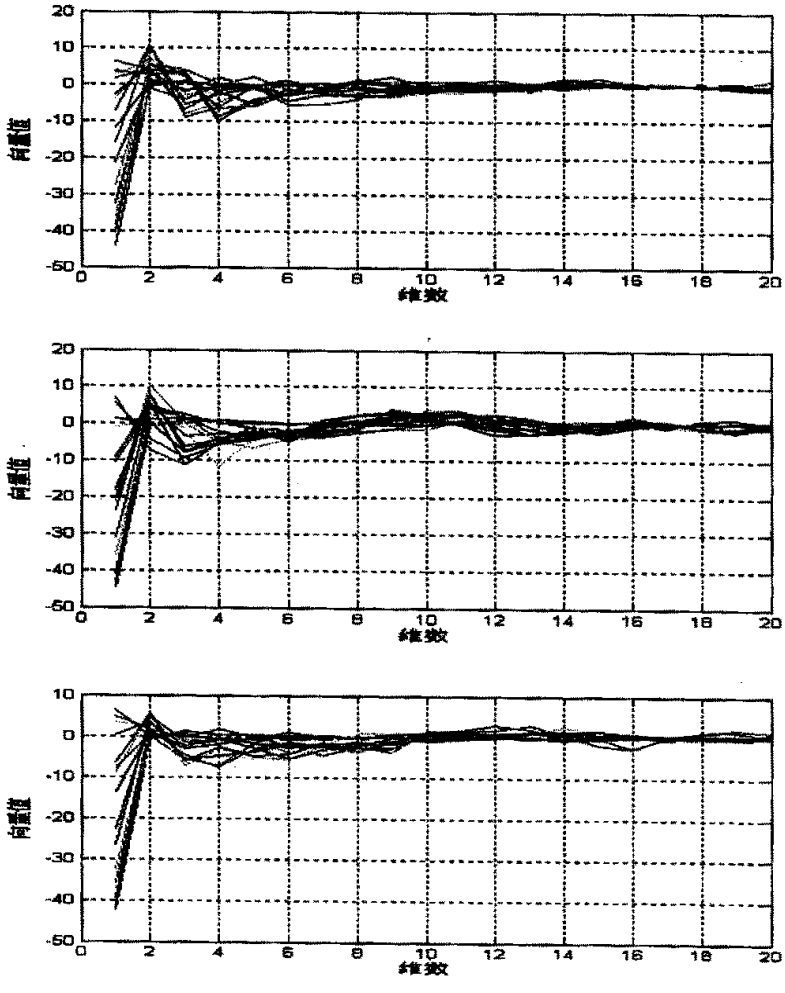
表 3.2 加噪前后识别率的对比

	第 1 组	第 2 组	第 3 组	第 4 组	平均
加噪前	92%	94%	90%	90%	91.5%
加噪（方差 0.0001）	86%	90%	86%	84%	86.5%
加噪（方差 0.0025）	65%	70%	69	60%	68.5%

表 3.2 反映了加噪前后识别率的变化。原始语音和加噪语音各作了 4 组实验，每组中说话人选取不同的短语进行测试，每个语句重复 50 遍，采样语音长度 2 秒。在噪声方差达到 0.05 时，识别率已经相当的低。

3.5.2 不同参数下的比较

为了得到最佳的系统运行参数，本文对音频采样率、计算倒谱时的帧移等几个参数对系统的影响进行了比较。图 3.14 显示了四个采样率为 6KHz 的语音样本计算得到的码字，而较高的采样率(大于 12KHz)使计算量成倍增加，识别率不会有什么变化，而现今普遍应用的 8KHz 电话质量语音，是说话人识别研究中经常采用的语音源。与图 3.13 中的码字略有不同，它对识别率的影响较噪声要小，当采样率低于 6KHz 时，识别率才会有明显的下降。



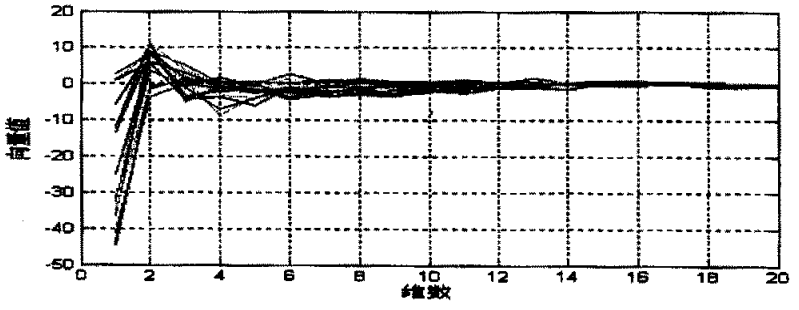


图 3.14 采用 6KHz 采样率时的说话人码本

对倒谱计算中的帧移参数，本文比较了帧长度为 256(对应 8KHz 采样率下的帧时长 32ms)，帧移为 64－160 之间的情况。帧移较小时，由于相同长度的语音数据会得到更多的语音帧，计算量会有较大增长。表 3.2 显示了不同帧移时不同说话人的码字距离，可以看出，码字距离的变化并不十分明显。结果表明，不同的帧移对识别结果的影响较小，考虑到 DSP 实现时的计算延迟，本文取帧移为 100。

表 3.3 不同帧移时不同说话人的码字距离

帧移	说话人 1	说话人 2	说话人 3	说话人 4	说话人 5	说话人 6	说话人 7
64	4.7977	4.5479	5.4155	5.4031	4.5409	4.0215	5.5026
100	5.0207	4.5448	5.5582	5.3434	4.7452	3.8495	5.7696
128	4.9877	4.5471	5.6741	5.3896	4.6607	3.9917	5.8100
160	5.1415	4.5798	5.5158	5.3952	4.6188	3.9913	5.2726

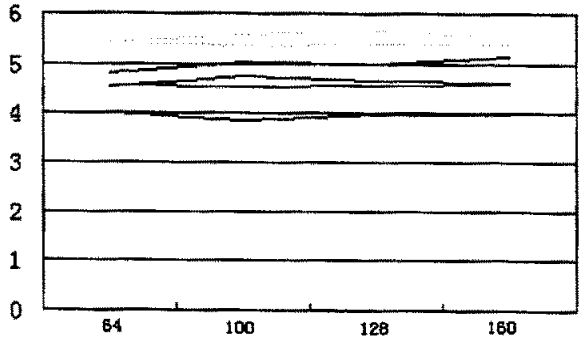


图 3.15 不同帧移时不同说话人的码字距离变化曲线

由表 3.5 可看出，增大聚类数，拉大说话人间的差别，也可有效提高确认效果。聚类数增大一倍，FA 和 FR 分别下降 5 和 6 个百分点。但不能无限地增大聚类数，只能根据实验结果适当地选择合适的类数。因当训练数据量有限时，一味增大聚类数必将造成每一类中训练向量数的减少，这将给模型的估计造成很大的误差，从而直接影响确认效果；聚类数也不能太小，因为类数小的时候，可能使不同类的向量被错误地分到同一个类别中从而造成误识。另外，聚类数的增加也相应增加了每个说话人的码字的存储空间。

表 3.4 采用不同 MFCC 阶数的结果比较

聚类矢心数目	集内总确认次数	错误拒绝次数	错误拒绝率	集外总确认次数	错误接受次数	错误接受率
16	100	6	6%	60	3	5.0%
32	100	5	5%	60	2	3.3%

表 3.5 采用不同聚类数目的结果比较

聚类矢心数目	集内总确认次数	错误拒绝次数	错误拒绝率	集外总确认次数	错误接受次数	错误接受率
8	120	15	12.5%	60	6	10.0%
16	120	9	7.5%	60	2	3.3%

3.5.3 测试语音长度的影响

本文实验研究了测试时长与说话人识别率的关系。训练语音经过预处理后提取出 20 阶 Mel 倒谱参数，作了 5 组实验，每组对每个用户选取不同的语句进行测试。实验数据见表 3.3。从实验数据可以看出，当测试时长增加时，系统平均识别率升高，语音测试时长在 1 秒到 3 秒时，识别率随测试时长的增加上升很快，3 秒以后升高的幅度趋向平缓。

在 DSP 系统对语音数据的处理中，由于语音数据所占的存储空间比较大，所以在之后的 DSP 实现及调试阶段，都必须根据可用的存储空间对语音长度作一定的限制。

表 3.6 不同测试语音时长对应的识别率

测试时长	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	平均
1s	62	78	81	83	73	75
2s	84	88	90	91	90	89
3s	90	91	94	96	94	93
4s	94	95	95	97	96	95

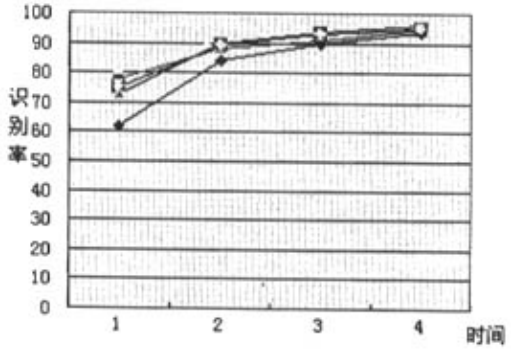


图 3.16 不同测试语音时长对应的识别率变化曲线

第四章 系统实现及优化

4.1 系统实现的硬件平台

本文是基于 TI 的 TMS320VC5402[40,41]数字信号处理器(简称 C5402)完成的。下面对硬件结构简要说明。

4.1.1 TMS320C54X 概况

TMS320C54x 系列的硬件由以下模块组成：

一、中央处理单元（CPU）

C54x 系列所有芯片 CPU 都相同，可以进行高速并行计算和逻辑处理。CPU 包含下列单元：

1. 40 位算数逻辑单元（ALU），包括一个 40 位的桶式移位器和两个独立的 40 位累加器。

2. 17 位乘 17 位并行乘法器和一个 40 位专用的加法器，用于非流水线的单周期乘法/累加操作。

3. 比较、选择、存储单元（CSSU），用于维特比算子的加法和比较选择。指数编码器，用来在一个单周期内计算一个 40 位累加器中数值的指数。

4. 两个地址产生器，包括八个辅助寄存器和两个辅助寄存器算术单元。

二、内部总线结构

C54x 有八条 16 位总线，包括四条程序/数据总线和四条地址总线，可以在每个指令周期内产生两个数据存储地址，大大提高了并行数据处理的速度。

三、特殊功能寄存器

C54x 共有 26 个特殊功能寄存器，用于对片内各功能模块进行控制、访问和其他管理。这些寄存器位于一个具有特殊功能的 CPU 映射存储区内。C5402 的特殊功能寄存器映射到在片上 DRAM 的 00－1A 单元。

四、存贮器

存贮区分为 RAM 和 ROM。RAM 又分为 DRAM(每个指令周期内进行两次

存取操作)和 SRAM(每个指令周期进行一次存储操作)。DRAM 除了进行双操作，还可以当作单操作使用。C5402 片内有 16K 双访问随机存取存储器（DRAM)。存贮区是存贮程序和数据的，既可以分布在片内，又可以在片外。一般，程序空间定位在 ROM 中，数据空间定位在 RAM 中。但数据空间不一定只定位在 RAM 中，也可以在 ROM 中。通过不同配置可以将数据空间映射到 ROM 中。

程序空间也不一定只在 ROM 中，也定位在 RAM 中。当运行程序时，可以用自动加载的方法将程序载入片内快速 DRAM 中，提高运行效率。

不同 C54x 的 ROM 容量有不同配置。不同 C54x 的 DRAM 和 SRAM 的容量、存储速度不同。 C5402 片内有 16K 的 DRAM，4K 已经固化的 ROM。

C5402 最大可以访问 64K 字的数据空间，可以访问 1M 的程序空间。除了可以访问数据空间与程序空间外，C54x 还可以管理 64K 字 I/O 空间。

五、片内外设

C54x 的片内外设依据型号各有不同。C5402 的外设具有如下特点：

(1) 主机通信接口(HPI)

HPI 提供 C54x 与主处理器之间通信的并行接口，实质是通过 C54xDSP 的片内存储器实现 C54x 和主处理器之间的数据交换。不同型号的 C54x 的 HPI 功能和配置不同。

(2) 串行接口

C54x 的串行接口随器件的不同而不同，共有四种不同的串行接口：同步串行接口，带缓冲的同步串行接口，时分复用串行接口和多通道缓冲串行接口(MCBSP)。C5402 有两个多通道缓冲串行口。

(3) 定时器

C5402 带有两个 16 位定时器。定时器可以有一个专门的状态位编程实现停止、重启动、复位和禁止。

(4) 直接存储器访问控制器(DMA)

DMA 控制器不需要 CPU 的参与，完成存储器映射区之间的数据传输。DMA 具有 6 个互相独立可编程的传输通道，允许有 6 种不同内容的 DMA 传输。

六、中断

C54x 具有丰富的中断系统，最高中断深度达九级。中断分为不可屏蔽中断与可屏蔽中断。可屏蔽中断又有硬中断，软中断。与 TMS320C54x 的硬件相适应，C54x 的软件是其一大特色。C54x 的软件是为信号处理专门设计的。C54x 具

有丰富的指令集和灵活的寻址方式。其中，有六条流水线操作，有硬件中断可进行九级中断，而且大部分中断可以通过软件灵活的控制。C54x 的中断可以由硬件驱动或软件驱动。C54x 系列 DSP 为用户构建系统提供了灵活丰富的中断资源。

4.1.2 说话人识别的系统平台

本说话人识别系统的开发平台是TMS320VC5402DSK开发板，下面简要说明其基本结构信息：

TMS320VC5402工作频率为最高100MHz

JTAG仿真接口——支持内嵌的和外部的JTAG仿真

控制接口——复位设备并提供外部中断

HPI——提供主机控制的、双向的PC与DSP之间的传送

引导模式——允许用户根据运行环境来选择无需引导、ROM 引导和HPI引导等方法

EMIF——支持外部的SRAM、FLASH存储器、外设。EMIF 也连接到了扩展连接器，这样就可以允许访问扩展板的存储器。

定时器接口——不在板上直接使用，而是连接到了扩展连接器。

McBSP0——不在板上直接使用，而是连接到了扩展连接器。

McBSP1——连接麦克风/耳机音频接口

DSK 包括电源稳压器，它为DSP的3.3V I/O电压和1.8/2.5V核电压提供必要的电流。

外部存储器。DSK提供64K x 16bit words（128KB）的SRAM和256K*16 words(512KB)的FLASH。板上的外部的SRAM和FLASH都是工作在+3.3V 的。

外部程序存储器。外部程序存储器的可用大小是取决于OVLY位的设置和MP/MC跳线的设置。如果OVLY位＝0并且MP/MC#＝0，那么程序存储器的空间0x0000～0xEFFF（60K words）映射到外部存储器，是FLASH还是SRAM决定于控制寄存器的FLASHENB状态位。在上电状态，FLASHENB位的设置是为了允许从FLASH 引导。然后软件清除此位，使具有1 个等待状态的SRAM 使用这个相同的存储器空间。如果MP/MC＝0，那么0xF000～0xFFFF是保留给片内ROM 和中断矢量表，并且外部程序存储器在0页是不可用的，但是在其它页可以使用，这些取决于OVLY位的设置。

麦克风接口。音频接口为工业标准的3.5mm的连接器，连接一个连接麦克风（J5）的音频输入，音频输入是交流偶合的，包括1个固定增益为10dB的放大器实现单端到差分的转换（在此之前，连接到DSP的McBSP1上的TLC320AD50 对其进行数字化）

4.1.3 本系统中对硬件的配置

语音输入采样设置为15Bit模式，并且对输入信号进行6dB的增益，采样频率设置为8000Hz。

经过前期的代码编写，得到程序的程序段长度为0x3CE4，因此将片内的RAM（长度为0x3F80）映射为程序空间，片外的RAM映射到数据空间。

由于代码长度的限制，DSP中只完成识别流程，说话人码本的训练在PC机中完成，训练好的码本直接固化到程序中，随程序代码一起烧写到板载的Flash存储器。

通过开发板上的3个LED灯作为系统的输出，输出信号包括：系统初始化完成，等待语音输入提示；测试者输入语音后，系统输出识别结果，3个LED代表3位二进制数，因此有效的说话人为0～6，如果输出7则表示说话人不在预置的数据库中。由于系统的存储空间有限，在目前的测试运行中，只预置了7个说话人的码本。

4.2 DSP 程序流程设计

软件算法开发较快，容易修改，灵活性好，但执行速度相对较低。相比之下，硬件的设计时间较长，修改困难，但运行速度较高。因此，对于一个实时系统平台的设计通常要用到折衷手段，这包括算法空间折衷、软件一硬件折衷、软件空间一时间折衷、硬件空间一时间折衷和算法一硬件结构之间的折衷。

在说话人识别系统平台中，DSP 可完成语音数据获取、特征提取、训练和识别等功能。语音数据可以从 DSP 平台的输入接口直接获取或通过主机调试口获取，然后由 DSP 进行实时处理或保存至主机硬盘中。训练阶段，在 PC 机中对说话人的语音数据进行特征提取后进行训练，生成的说话人参考模型参数（码本）保存到开发板板载的 FLASH 中；识别阶段，对测试语音数据进行特征提取后用说话人参考模型进行模式匹配，得出识别结果。

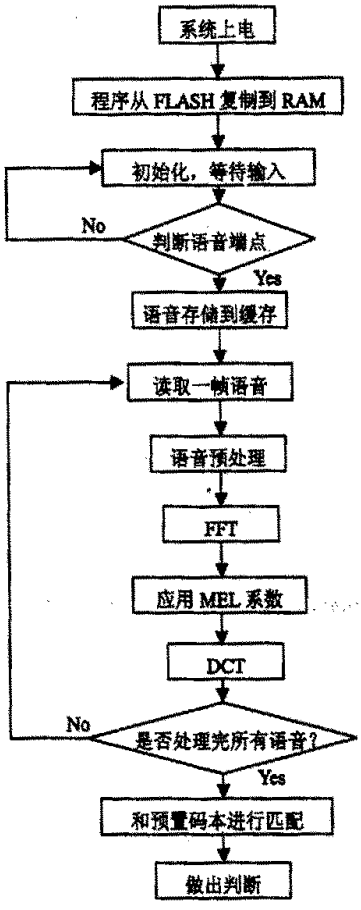


图 4.1 识别过程的 DSP 流程图

采用语音信号的采样率为 8000Hz，采样后的语音信号通过预加重运算、分帧、最后将每帧语音转换成相应的特征矢量，本系统的特征矢量为 20 阶 MFCC 系数，对应于一个短时帧的 20 阶 MFCC 系数组成的 20 维矢量空间的一个特征矢量。

由 3.5.3 中的试验结果可知，测试时长的增加带来识别率的提升，但同时也增加了识别的计算量，语音长度为 1 秒时，缓存语音需要的存储空间是 8K 字，语音长度为 4 秒时，缓存语音需要的存储空间是 32K 字。这里设置采样的时间长度为 2 秒。

然后是对输入特征矢量进行矢量量化，即用 LBG 算法对话者的训练，集中

所有的训练矢量（帧）训练出 VQ 码本。这里的关键是 VQ 码本容量的选值。实验结果表明，当码本容量小于 64 时，随着码本容量增加，正确识别率仍有提高，当容量大于 256 时，提高就不明显，故目前常用的容量值取为 64、128、256，在本系统中由于系统存储空间有限，取值为 64[37]。

4.3 软件优化

为了适应 DSP 处理器存储空间小和运行时钟速度较低的特点，对程序进行了部分的优化，主要目的是提高运行速度。

对 FFT 算法进行优化。FFT 有两个可行版本，一是 C 语言实现，另一个是 Ti 公司的 DspLib 库中的实现，前者使用浮点数据类型，计算精度和 PC 机基本一致，而后者由于是定点计算，结果存在一定误差，但是执行速度比前者快很多。在程序实现的早期，采用的是 C 语言实现的算法，主要是考虑到要将中间结果和 Matlab 的仿真结果对比，代码调试通过后，再将这部分改为定点方式，经过验证，最终计算结果的误差对识别结果的影响可以忽略。

预加重。由于在预加重时使用一阶高通滤波器，这与 DSP 流水线工作相冲突，因此需强制每个采样值为两个机器周期。

4.4 系统测试及相关结果

在仿真阶段，本系统在 PC 机上对 7 人（2 名女生、5 名男生）进行训练语音和识别语音采集录制。对‘123’、‘abc’两个短语进行 20 遍采集，其中 10 遍用于训练码本，10 遍用于测试，识别正确的概率达到 94%。

DSP 实时测试时，码本采用仿真阶段存储的对应说话人的码本。在系统启动后，通过开发板上的 3 个 LED 灯提示系统初始化完毕，等待语音输入，测试者说话后，系统经过 6 秒左右的时间输出识别结果，3 个 LED 代表 3 位 2 进制数，因此有效的说话人为 0～6，如果输出 7 则表示说话人不在预置的数据库中。实际测试中，同一个测试者按照同样的语速语调重复同一短语 20 次，被正确识别的概率大于 92%，比仿真阶段有所降低，主要原因是由于语音采样长度受存储空间限制而减少。

与文献[37][43][44]中的研究结果比较，采用多算法组合识别方案的研究中，识别率已经可以达到 97%，本论文的结果在识别率方面并不占优势，但考虑

到系统的实现平台为 DSP，不可能采用复杂的组合设计方案，为达到实时性的要求，采用本文的设计方案。而使用相似方案的文献[49][17]中识别率指标大致在 89%到 94%之间，本文的结果显示在识别率方面有所改进。

系统的开发平台如图 4.2 所示。

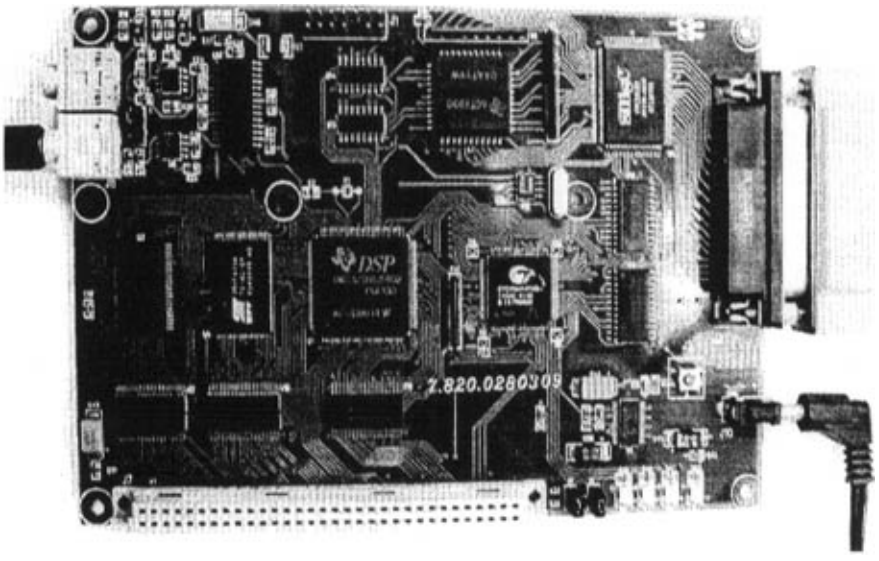


图 4.2 系统开发平台图

第五章 总结与展望

5.1 论文总结

本文回顾了说话人识别技术发展的历史，阐述了特征提取、模式匹配和判决规则等说话人识别中的主要技术理论。详细地讨论了基音频率、线性预测系数及美尔倒谱系数等特征提取方法，以及动态时间规整、矢量量化、隐马尔可夫模型等模式匹配算法的原理及实现流程。

本文设计实现了一个基于 DSP 的说话人识别系统，采用美尔倒谱系数作为特征参数，以矢量量化作为匹配算法，对系统在不同参数下的识别性能进行了仿真，选取最优的方案在 TMS320 C5402DSK 上实现了说话人识别系统。论文针对系统的实现平台上的实际应用进行系统设计、提高系统的识别率、可靠性和减少识别时间。经过测试，系统运行正确，达到预期目标。

论文的主要工作是：

1.研究了说话人识别的基础理论，对几种常用算法的复杂性和效率进行了比较。

2.使用 Matlab 和 Visual C++对系统算法进行了仿真。

3.采用特征提取算法 MFCC 和模式匹配算法 VQ 的组合，在 TMS320 C5402DSK 开发板系统上实现了说话人识别系统。

4.优化了 DSP 算法的逻辑结构，使得系统的存储空间利用率和时间效率得以提高，程序占用空间从 0x43f1 减少到 0x3ce4，其中.text 段长度从 0x3190 减少到 0x21ff，程序运行中，从输入语音完毕到显示结果的时间，也从 11 秒减少到 5 秒。

5.2 待研究的问题

由于时间和其他一些原因，本次设计还有一些工作需要继续完善，系统性能有待提高。在后续的工作中可将重点放在以下方面：

(1) 在基于 VQ 的说话人识别算法中，特征矢量可以有多种选择，今后我们还可以针对不同的组合参数进行实验，以对算法进行改进。同时，我们可以寻找更有实用价值、更适合 DSP 实现的算法。

(2) TMSC5402DSK 开发板上采用 FLASH 来存储程序和码本，由于 FLASH 只有整片擦除和块擦除方式，这给码本更新过程带来不便，可以在系统上扩充一片掉电不丢失的 SRAM 来代替，这样就可以对码本进行选择性的更新。

(3) 该实验系统的整体运行有时不是很稳定，并且还没有达到完全实时化的要求。今后需要进一步改进硬件系统，同时进一步优化软件设计。

阻碍说话人识别系统实用化的最大障碍仍然是系统的识别性能问题，即对语速快慢、语调高低的适应性以及间隔一段时间后系统识别性能的稳定性。如果想进一步解决该问题，建议重点研究语音的特征提取问题。已有的语音特征的潜力已经挖掘得差不多了，所以要注意研究新的语音特征。

致 谢

本论文是在导师史燕副教授的严格要求和精心指导下完成的。从论文的构思、开题，到论文的每一细节部分都凝聚着导师的心血。在两年多的研究生学习期间，导师以其严谨的治学风格、渊博的学术知识和积极创新的生活态度，给予我莫大的教诲和启迪。

感谢范俊波老师给我的指点与帮助。

感谢赵迎辉、郑文生、顾成威同学在我论文期间的帮助与支持。

曾海涛

2006-5-5

参 考 文 献

[1] 易克初，田斌，付强《语音信号处理》国防工业出版社，2001年6月

[2] 赵力《语音信号处理》机械工业出版社，2003年3月

[3] 朱民雄，闻新，黄健群，周露《计算机语音技术》北京航空航天大学出版社，2002 年1月

[4] 谷荻隆嗣《语音与图像的数字信号处理》科学出版社，2003年9月

[5] 胡光锐《语音处理与识别》上海科学技术出版社，1994年

[6] 杨行俊，迟惠生等《语音信号数字处理》电子工业出版社，1995年

[7] L.Rabiner, B.Juang《Fundamentals of speech recognition》Prentice Hall，1993年

[8] B.Gold, L. R.Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain", J.Acoust.Soc.Am, Vol.46,pp. 442-448, 1969

[9] B.S.Atal, "Automatic speaker recognition based on pitch contours", J.Acoust. Soc.Am, Vol.52, 1972：1687-1697

[10]Goutam Saha, Sandipan Chakroborty, Suman Senapati, An F-Ratio Based Optimization Technique for Automatic Speaker Recognition System, IEEE INDIA ANNUAL CONFERENCE 2004. INDIMN 2004：70-73

[11] Joseph P.Campbell, JR Speaker Recognition： A Tutorial [J]. Procedings of the IEEE, Vol 77,No.9, 1997; 1437-1462

[12] 杨行峻、迟惠生.语音信号处理.北京：电子工业出版社，1995

[13].岳喜才，伍晓宇.用神经阵列网络进行文本无关的说话人识别.声学学报，2000,25(3)：230-234

[14].William M. Campbell, Khaled T. Assaleh, Charles C. Broun.Speaker Recognition With Polynomial Classifiers. IEEE Trans On Speech and Audio Processing. 2002,10(4)：205-212

[15] Chularat Tanprasert, Varin Achariyakalporn. Comparative study of GMM, DTW, and ANN on Thai speaker identification system. In： ICSLP,Beijing, 2000：718-721

[16] Chih-Chien Thomas CHEN, Chin-Ta CHEN, and Shung-Yung LUNG. Efficient Genetic Algorithm of Codebook Design for Text Independent Speaker Recognition. IEICF Trans Fundamentals, Vol.E85-A,No. 1 1,2002：2529-2531

[17] 张炜，胡起秀，吴文虎.距离加权矢量量化文本无关的说话人识别.清华大学学报(自然科学版),1997,37(3)：20-23

[18] J.Oglesby and J.S.Mason, "Radial basis function networks for speaker recognition", Proc. IEEE ICASSP, pp. 393-396, 1991

[19] Georger Doddington, Speaker Recognition - Identifying People by Their Voice, Procceedings Of The IEEE, Vol. 73, No. 11, Nov, 1985, 1651~1663

[20] Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's, Tomoko Matsui and Sadaoki Furui, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 2, NO. 3, JULY 1994

[21] 荆嘉敏 刘加 刘润生 基于HMM的语音识别技术在嵌入式系统中的应用，电子技术应用，2003年第10期

[22] HASSEN SEDDIK, AMEL RAHMOUNI and MOUNIR SAYADI, Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier, IEEE, 2004

[23] CHERIF Adntne, PITCH AND FORMANTS EXTRACTIONALGORITHM FOR SPEECH PROCESSING, Electronics, Circuits and Systems, 2000. ICECS 2000. The 7th IEEE International Conference on Volume 1, 17-20 Dec. 2000 Page(s):595 - 598 vol.1

[24] Rabiner et al., "A Comparative Performance-Study of Several Pitch Detection Algorithms," IEEE Trans. ASSP, Vol. ASSP-24, No. 5, October 1976.

[25] 朱建新 杨小虎 叶荣华,生物认证系统性能评估研究,计算机工程与应用,2002.16,Page:66-68

[26]Zhu Jianxin,Yang Xiaohu,Dong jinxiang, Fingerprind-based Authentication in Network Environment[C].In:ICYCS' 2001,2001

[27]Anil Jain,Lin Hong,Sharath Pankanti. Biometric Identification [J],COMMUNICATIONS OF THE ACM, 2000:43, Page:90-98

[28] Mueen F. Ahmed A.Sanaullah Gaba.A, Speaker recognition using artificial neural networks, Students Conference, ISCON'02. Proceedings IEEE Volume 1, 16-17 Aug. 2002 Page(s):99 - 102 vol.1

[29] 荆嘉敏 刘加 刘润生,基于HMM的语音识别技术在嵌入式系统中的应用，《电子技术应用》2003年第10期,page:12-14

[30] Bernd Burehard, Ronald Romer, Oliver Fox. A Single Chip Phoneme Based HMM Speech Recognition System For Consumer Application, IEEE Transactions on Consumer

Electronics, 2000;46(3)

[31] Speaker recognition: a tutorial . Campbell, J.P, Jr. Proceedings of the IEEE, Volume: 85 Issue: 9, Sept. 1997 Page(s):1437 — 1462

[32] 宁飞 陈频，说话人识别的几种方法，电声技术，No.12,2001

[33] 李宵寒 戴蓓倩等，高阶 MFCC 的话者识别性能及其噪声鲁棒性，信号处理，Vol.17, No.2, April. 2001

[34] Furui. S., Cepstral analysis technique for automatic speaker verification. Speech, and Signal Processing[see also IEEE Transactions on Signal Processing], Issue: 2, Apr 1981 Page(s): 254 -272

[35] An introduction to speech and speaker recognition. Peacocke, R.D.; Graf, D.H. Computer, Volume: 23 Issue: 8, Aug 1990 Page(s): 26 -33

[36] Li Liu; Jialong He; Palm, G., Signal modeling for speaker identification, Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on Volume 2, 7-10 May 1996 Page(s):665 - 668 vol. 2

[37] Ezzaidi, H.; Rouat, J., Pitch and MFCC dependent GMM models for speaker identification systems, Electrical and Computer Engineering, 2004. Canadian Conference on Volume 1, 2-5 May 2004 Page(s):43 - 46 Vol.1

[38] Guiwen Ou; Dengfeng Ke, Text-independent speaker verification based on relation of MFCC components, Chinese Spoken Language Processing, 2004 International Symposium on, 15-18 Dec. 2004 Page(s):57 - 60

[39] Kim, S.; Eriksson, T.; Hong-Goo Kang; Dae Hee Youn, A pitch synchronous feature extraction method for speaker recognition, Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on Volume 1, 17-21 May 2004 Page(s):I - 405-8 vol.1

[40] TMS320VC5402 FIXED POINT DIGITAL SIGNAL PROCESSOR, www.ti.com, SPRS079E - OCTOBER 1998 - REVISED AUGUST 2000

[41] TMS320VC5402DSK 中文用户手册

[42] Soong, F.; Rosenberg, A.; Rabiner, L.; Juang, B., A vector quantization approach to speaker recognition, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85. Volume 10, Apr 1985 Page(s):387 - 390

[43] Kevin R. Farrell, Richard J. Mammone, and Khaled T. Assaleh, Speaker

Recognition Using Neural Networks and Conventional Classifiers, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 2, NO. 1, PART 11, JANUARY 1994

[44] Jialong He, Li Liu, and Gunther Palm, A NEW CODEBOOK TRAINING ALGORITHM FOR VQ-BASED SPEAKER RECOGNITION, IEEE, 1997, page:1091-1094

[45] 朱建新 杨小虎 董金祥 曹哲新，生物认证系统性能评估研究，计算机工程与应用 2002.10 pp:92-95

[46] Sadaoki Furui, Vector-Quantization-Based Speech Recognition and Speaker Recognition Techniques, IEEE, 1991, pp:954-957

[47] 丁佩律 张立明，结合主分量分析及 Fisher 准则的说话人识别方法研究，电路与系统学报，第 7 卷，第 1 期，2002.3

[48] 李虎生 刘 加 刘润生，语音识别说话人自适应研究现状及发展趋势，电子学报，2003

[49] 孙林慧 叶蕾 杨震，说话人识别中测试时长与识别率关系研究，计算机仿真，2005 年 5 月

[50] D A Reyolds , R C Rose. Robust, Text Independent Speaker Identification Using Gaussian Mixture Speaker Models [J]. IEEE trans. On Speech and Audio Processing , 1995 , 3 :72 - 83.

攻读硕士学位期间发表的论文

[1] 曾海涛，杨光. 风洞运动机构振动监测系统. 电子测量技术. 2005 年增刊（2005 全国虚拟仪器学术交流大会论文集）