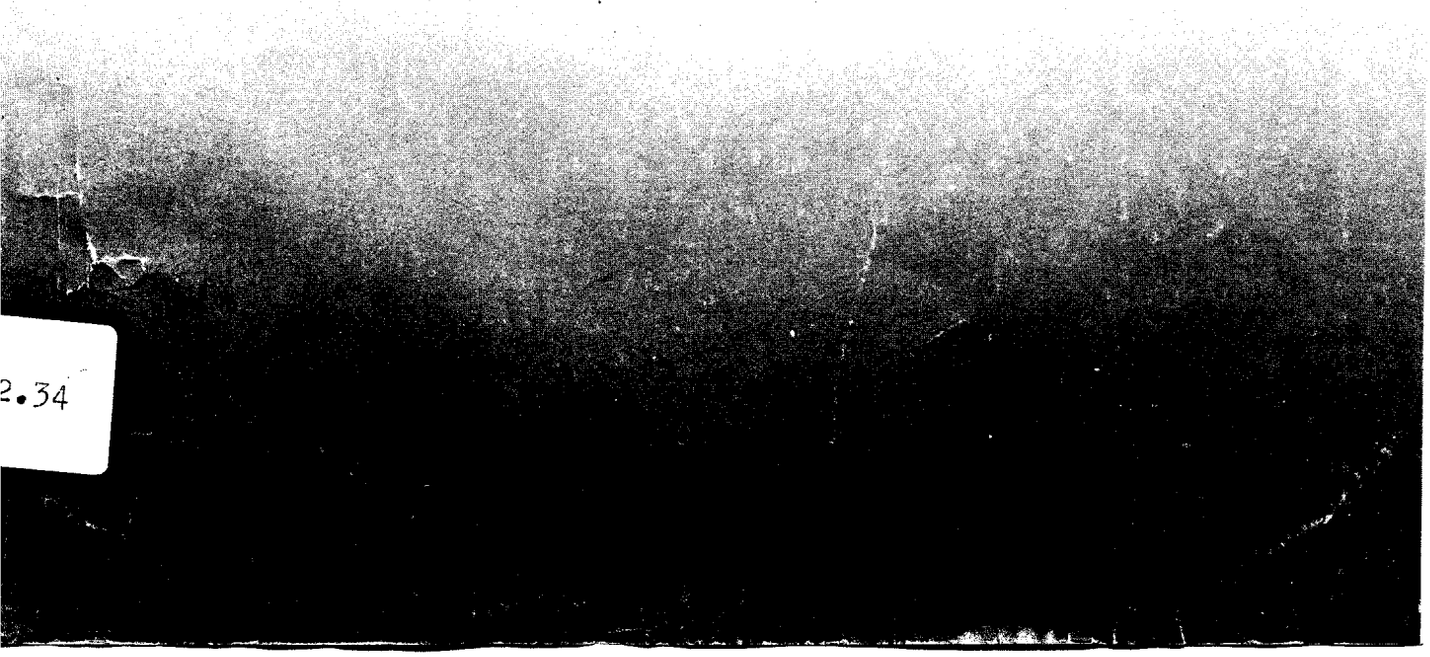
近代语音识别

陈尚勤　罗承烈　杨　雪



内 容 提 要

“语音识别”主要指用机器在各种环境下识别和了解语音和其他声音，从而根据其信息执行人的各种意图，是近十几年发展起来的新兴学科，它在计算机、信息处理、通信及电子系统、自动控制等领域中，工业、军事、交通、医学、民用诸方面有着广泛的应用。

本书在编著中既着重基本理论、思路和方法的阐述，又着重近期出现的有价值的思想和方法以及在工程设计等方面的实际应用。本书列有大量有关参考文献和有实用价值的试验结果、电路结构及程序。

本书适用于高等学校研究生、高年级学生的教学及各种研究、设计单位工程技术人员参考之用，也宜作为信息等学科科技人员和教学人员知识更新的参考读物。

近代语音识别

陈尚勤　罗承烈　杨　雪

*

电子科技大学出版社出版
（中国成都建设北路二段四号）
电子科技大学出版社印刷厂印刷
四川省新华书店经销

*

开本 787×1092 1/16　印张 16.875　版面字数 425千字
版次 1991年3月第一版　印次 1991年3月第一次印刷
印数 1—5200册
中国标准书号 ISBN 7-81016-260-8/TP·19
（15452·117）　定价：5.80元

绪　言

“语音识别”主要指用机器在各种情况下，有效地了解、识别语音和其他声音，从而根据其信息执行人的各种意图，是近十几年发展起来的有理论和实用价值的新兴学科。从计算机大学科角度来看，它可视为智能计算机的智能接口科学；从信息处理大学科来看，它属于信息识别的一个重要分支；从通信及电子系统、电路、信号及系统等大学科来看，它又可视为信息或通信系统的信源处理科学；而从自动控制大学科来看，它则可看成是模式识别中的一个主要部分。另一方面，用计算机等机器来识别人的语音，又使语音学家及生理学家也感到兴趣。因此，国内外高等学校逐步开设有关语音识别课程，很多研究单位对语音识别的理论及实际应用作了广泛的研究。目前，国内外有关论文每年达数千篇之多。我们近八、九年来也在这方面进行了研究生和本科高年级学生的教学工作和科研工作，深感若能将有关此学科散处各处的有价值的资料，包括基本理论、方法、新的设计思想等加以消化、归纳、整理，结合我们的教学体会和一些科研成果（如高抗扰语音识别控制系统，仿人型机器人、视听觉语言行动智能机器人等）予以系统地编纂成书，可能对有关的教师、学生和研究设计工作者有所裨益，这就是我们编著“近代语音识别”一书的目的。

我们编著此书的主导思想是：

（1）内容包括基本理论、思路和手法，又着重近代出现的有价值的思想与方法。由于取材不少来自近期的文章、杂志，为了使读者阅读时便于与以往知识衔接和作进一步的考查，在书末列出了各章主要的有关参考文献名称；

（2）内容着重理论、方法的阐述，也注意便于读者实际应用，所以书中包含工程设计的方略、步骤、有代表性及实用价值的试验结果和有关的电路结构及程序等。

书中的不少思想来自与我们多年共事的学友或他们的启示。他们中不少是当年从事此项研究的研究生。在此我们谨向李旭、王新扬、张世平、田莉、梁虹、张鸿、黄跃新、周红梅、童赛美、杜笑平、张福洪、骆安、冯军、李立忠等同志以及刘亚康、魏鸿骏、邱致君、汪亚南等老师衷心致谢。

本书承四川大学杨家沅教授、机械电子工业部第十研究所姚居济研究员及电子科技大学李在铭教授审阅。尤其是杨家沅教授作了逐字逐句的审校，并提出了宝贵的意见，我们在此诚恳地表示敬意和谢意。

由于我们水平有限，书中定有欠妥之处，敬希读者予以指正。

编　著　者

一九九〇年春

目　录

第一章　语音识别的目的与内容概述

1.1　语音识别的目的及内容安排总体思想

语音识别装置现指用机器设备(主体为计算机)来识别人所发的语音，按其含义照其意旨执行命令，例如实时执行军事指挥员或飞机舱中驾驶兼战斗员的口述命令(如突然操纵，飞机技巧操作行进中的实时发炮)，汽车驾驶员口述实时紧急刹车，工厂中工作者口述命令“声控机器人”作实时危险操作，计算机使用者口述指令或程序操纵计算机，在人手已被占用的各种场合下声控(尤其是遥控)各种机器操作，以及作为各种语种翻译机或打字机的口述输入设备和声控印刷排版等。总之它是人通过语音使机器了解人的意图从而使之执行命令的设备。

利用语音使设备了解人脑反映较之利用书写文字(图片)或按钮有如下的优点：

(1) 人脑的意图反映到发出声音较反应到输入文字或按钮操作要迅速得多，可达毫秒量级；

(2) 将语音信号传送到语音识别机可在黑暗中进行而此时图片文字或按钮等信息传送方式则较难做到；

(3) 语音向识别机传送信息一般无严格的方向限制，这点对于图片文字传送则较困难。

因此语音识别设备有重要应用价值，各方面对此日益提出需求。另一方面，由于计算机及计算技术，模式识别和信息处理技术，声学技术等的发展使满足各种性能需要的语音识别装置的实现也成为可行。在此形势下，有关各种语音识别技术和系统的开发、创造、设计的新思想、新方法和新成果不断出现，急需人们作总结、归纳、整理或有目的的补充和创新，使之有利于用于实际，为社会服务。本书的编写即是拟在这方面作一些工作。

还需说明的是，有关语音识别的技术，如特征选取和提取的方法，学习、识别方法(包括发同一音各次时间长短不一致的处理方法)，对抗干扰和噪声的方法，多个作为标准用的模板数据的合理合并以压缩所需存储量以及加快处理速度的思想和技术等都不限用于人的语音的识别。有些直接或稍加修改便可用于其它声音或振动波形的识别。就是说语音识别技术可以扩充应用到多方面，例如地震波性质识别，利用机械声波识别来判别机械故障，利用声波特征来判别同轨火车相撞的危险性，利用心电、脑电图波，脉搏波的识别作疾病诊断等。此外，语音识别技术中的一些思想和方法由于与除振动波识别外的其它模式识别有一定的共通性，所以也可能推广应用到这些领域(如图形识别)中。

内容安排的总体思想

就语音识别机的应用场合而言，可以分为：(1)弧立词(命令词)语音识别机；(2)连续语音识别机；(3)话者识别机等。

就语音识别的重要组成部分及所需解决的重要问题而言有：特征提取部分；学习识别部分；重要的预处理部分和抗噪声干扰问题；发音时间不一致的问题；当同一字音各训练样本中含有大量特征有较大差异的数据时如何建立识别时用的模板的问题等。

在本书中：

（1）着重以语音识别机重要组成部分所采用的技术及语音识别领域中所需解决的主要问题来取材，编排叙述也按此进行，这样便于说明问题的实质及易于从所处理问题的思想方法中得到启示。

（2）另一方面，语音识别系统具体的设计及实时实现问题也是重要的，所以我们选择了一种系统在设计和实现方面作较全面的讨论，它即孤立词（命令词）识别系统（其它则作一般讨论）。由于此类系统本身用途甚广，且其有关知识对其它类型识别系统而言有通用性并易于推广，所以稍加补充一些知识即可用于其它类型系统（如在识别部分加用适当语义信息等则可用于连续音识别机等）。

（3）汉语的孤立词音及连续音均有其特点（主要是它宜于用“音节”为基本研究对象），它们致使特征的选取，字节的分割、动态时间匹配方法的选取等也具有特点。我们在各章的取材中特别注意内容包括汉语字（或词）音学习识别特征选取等有关的问题。第六章内容更是直接与汉语字音特征选取的实验研究有关，第三章中的四声调识别也主要适用于汉语音的识别。

（4）在取材中注意有价值的新思想、概念和方法，尤其着重其物理概念的阐述。

此外，书中还列出大量参考资料名称供读者进一步查阅。为了避免重复，突出要点，一般较成熟的内容，书中不作详述，读者可参阅注明的文献。内容具体安排见1.4节。

1.2 语音识别系统基本结构及语音发音模型

语音识别的基本结构如图1.1所示，其中各部分的功能综合起来可完成语音识别的基本功能。在学习时“特征提取”部分将数字语音信号按短时段（长度8～20毫秒，邻段可重叠）将数据分组（帧），由拟定的特征提取程序算出每短时段（帧）的特征向量（维数视需要而定），并将它们组成序列予以存储，称之为该字的模板（序列）。在计算中应存储有所有该系统需识别字的模板。

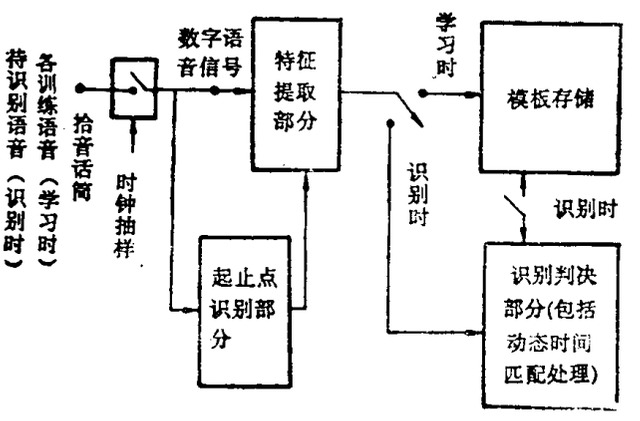


图 1 1

在识别时，输入语音信号经与学习时相同的特征提取部分得到待识别字的特征向量序列，再在“识别判决部分”中将输入的识别样本序列与从模板存储部分搜寻到的各字的模板序列分别作比较（即按规定的“距离”或“计分”的定义求出对各字而言的距离或计分），以距离最小（或计分最大）者判为所识别字。在求距离过程中应考虑到建立模板与识别时，发同一字（或词）的各瞬间速率可能有一定的差别，故在作比较计算识别信号与模板信号的总距离（计分）时，应使所规定的求总距离（计分）的方法在上述意义下合理。所用的方法有三大类。其一称为动态时轴弯曲（DTW）方法，即在时域中求识别样的某时段与模板样相应段的距离时，以其与准确时间对应的前后若干样之间的各种距离中之最小者定义为其距离（即应用了在

时域内的弹性匹配概念)。对于另一类，每个字的模板不直接以特征向量时间序列的方式存储，而是以态图的形式存储，图1.2所示的是一例子。图中 $a_{ij}(i=1\sim3, j=1\sim3)$为由 i到 j 态的转移概率，$b_i(i=1\sim3)$为 i 态输出各特征的概率。各a_{ij}、b_i是在学习过程中根据各训练样序列估算出来的。准则是对给定的所有训练序列，由始态(现为1态)到终态(现为3态)得到的输出概率之和最大。当 a_{ij} 及各 b_i 都是定值时，对于每一训练样可得到一输出概率，求此概率的方法可用最优路由问题中的最优路由求法及在此最优路由情况下的计分求法(即 Viterbi 算法)。此时学习的目的即是求出各态图模型中的 a_{ij}、b_i。在识别时，令识别样也通过每个字的态图模型，按 Viterbi 算法得到输出概率(即本法的计分)。以计分最大的态图相应的字为所判别字。由于某字的各训练样总起来包括了该字各种发音速率变化的状况(体现于模型中a_{ij}等参数的值)，故与DTW 法有异曲同工的效果。此类方法常称为 Markov 模型(MM)法或隐Markov 模型(HMM)法。第三类解决语音动态时间匹配的方法是先算出训练样帧序列中每个帧的“累计特征差”(如第3帧的累计特征差即由字音开始起第1、2帧的特征差加第2，3帧的特征差)，并只将累计特征差为所规定的 n 个(如8个)值的帧号及相应的特征予以存储(称之为 n 个关键帧)。在识别时按同法得到 n 个关键帧，与模板的各关键帧一一相应作比较，将 n 个距离相加得到与该模板间的总距离。同样得到与其它各模板间的总距离，而以其中最小者所相应的字为判别字。由于在本法中将累计特征值相同的帧对应求距离，故能保证识别样与模板样各相应的音素在求距离时基本对齐。因为特征常统称为谱，故此类方法常称为“声谱差”法。

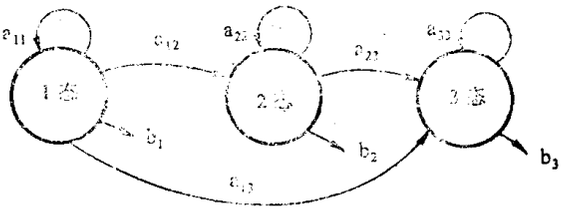


图 1.2

识别系统一个最根本的先决问题是合理地选用特征。准则是能使同字音间距离的区别小而异字音间距离的区别均值大。纯用数学计算方式按此准则选择特征并不容易，一般应借人发音的物理模型作启示，从中暂定一些特征，再经试验和计算逐步修改或调整来决定。人的基本发音电模型见图1.3。当发浊、清音时，浊/清音开关分别处于1、2端。时变数字滤波器模拟发音时声道的作用。在处理时，常认为在一个短时段内声道保持不变，故当假设时变滤波器的 Z 传输函数为全极点型，即

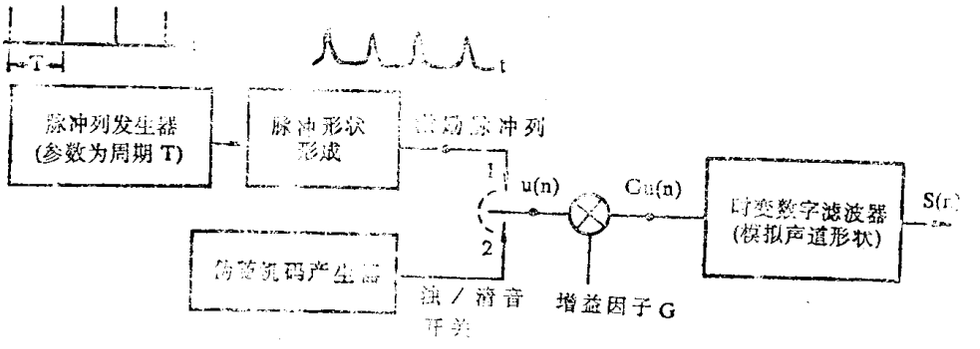


图 1.3

$$H(Z)=\frac{1}{1-a_1Z^{-1}-a_2Z^{-2}-\cdots-a_pZ^{-p}} \tag{1.1}$$

时，其中各系数 $a_i(i=1\cdots p)$随短时段而变动。a_i 等可由语音信号 $s(n)$求出，而作为其特征，

称为线性预测系数特征(LPC特征)，它们显然表征发音者在各短时段内声道的特性。同样，脉冲发生器的周期 T 代表声带振动的周期。G 代表发音的响度。总之若以各LPC、T、G为 $s(n)$的特征，则它们各具物理意义，彼此间的关联性较小，所以是一组有效的特征。从中我们可得启示，从物理模型出发，以其中的参数作为识别系统的特征，往往是有效的。当然，从另一方面看上述的各特征并非绝对独立，图1.3中电模型也并非完全代表发音模型，上述特征所需的计算量有些可能嫌大，所以对这些特征的选取还需经实践考验，在有些应用场合可能作适当的修改、补充或变形，成为另一些形式的特征，如倒谱特征，带通滤波器组(BPFG)输出特征，过零率特征，LPC误差特征等。

1.3 从实际需要提出的问题及解决问题的宏观思路与方法

具有上节所述基本结构的识别系统一般说来仅能完成基本的语音识别功能，但从实际出发，人们往往对识别系统的性能等提出了更多的要求，如

(一)鉴于目前较成熟的识别系统的识别本领尚远不如人耳，尤其体现在对于属于同字的各次发音，当人耳(包括发音及听音者耳)完全能识别为同字(且认为各次发音差别很小)时，在识别系统中所示出的差别有时颇大，即彼此间距离大而经常错判为不同字音。解决此问题思路之一是在识别机识别部分之前模仿人耳对输入音(或输入特征)的听觉处理，使其示出的对于同一字音的各发音间的距离缩小。这处理程序可以通过对动物耳的生理解剖分析或结合用改变有关参数考察距离计算值的试探法来拟定。然而为何当人多次发同一字音时，本人及他人从听觉来说非常一致而测得的各特征向量却常差别颇大呢？要得到理论依据还需进一步研究发音模型。经研究得到修正后的模型见图1.4。从图中可看出，它反映了人发音时会有信息反馈回脑，根据自己的听觉感受，随时部分调整发音电模型的有关参数，而这样长期以来形成习惯，只要本人听觉感受一致，即不去精确控制或调整上述各参数，以致它们可在与听觉机理有关的一定范围或规律下变动。图1.4模型启发了我们除研究发音机理外尚需研究听觉模型，并将此部分知识用于识别系统中特征提取的后处理部分中。事实证明这种作法可使识别率显著提高。

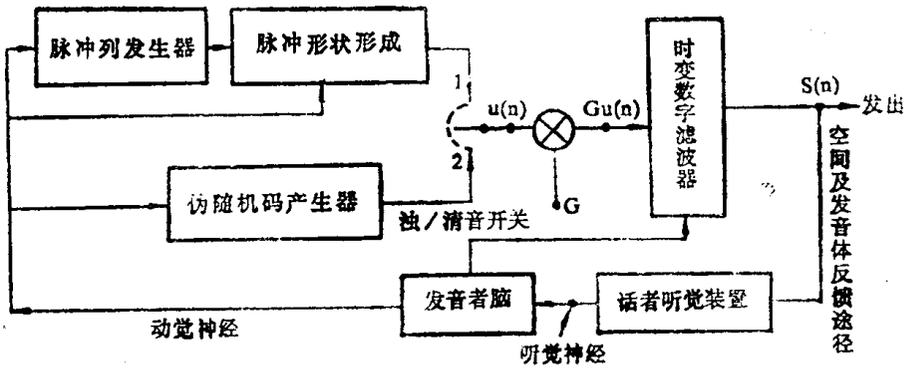


图 1.4

(二)人们常希望识别系统能为多人服务，即能识别多人的语音。使用者可为规定的一组人，他们事前共同训练识别机，或使用者为任意人，应用由具有代表性的一组人所训练的识别机。也有这样的应用方式，使用者可为任意人，他(她)应用前只要求发少量规定的训练音

使机器对此人适应（属自适应模板参数型多人系统）。解决此问题的方法是：利用特征向量（特征空间中的点）的适宜“聚类”技术；在DTW系统中多个模板时间序列的合理归并技术、平均技术；在HMM系统中如何利用多个输入信号序列的数据决定某字的HMM中输出概率(b_i)密度函数内参数的技术。在自适应模板参数型系统中，则采用这样的技术，即在原有模板参数的基础上，只用少量新添数据来修改参数，而能与用完整数据重新作全盘计算所得的基本相同。

（三）有些识别系统需在较严重的噪声和干扰（包括人声或机械声干扰等）的环境下工作（在战斗机舱中作声控命令指挥即为一例）。为了减少噪声、干扰影响，可从三条思路出发：

（1）建立消除、减少噪声和干扰的前端电路。如用自适应噪声对消，（此时设有拾取噪声样本的参考话筒），自适应信号增强，线性Weiner滤波，利用信号与噪声干扰的自、互相关差别的方法和直接适当地从总谱中减去噪声频谱的方法等。

（2）利用适当安排各话筒的位置及选用适当的话筒结构的方法。如用接触式话筒，将话筒置于氧气屏蔽罩内及合理设置和设计拾话音、拾噪声话筒间的相对位置及相应的处理电路等。

（3）用噪声补偿技术。即在DTW（或HWW）系统中作距离（或概率）计算时注意考察每个短时段中信号受噪声影响的程度，判知何者已受掩蔽，何者为基本上未受影响的有效段等，适当修改距离（或概率）定义，使之加重有效信号段测得的特征距离（概率）信息对总距离（概率）的贡献，减少受掩蔽段的贡献，从而减少噪声对识别性能的影响。

（四）人们往往希望进一步提高识别系统的识别率，或者保持高识别率及不过多增加计算机存储容量的情况下增加待识字表中的字数。解决此问题可从多方面考虑。如：

（1）通过对采用不同特征的系统作分析和对疑难识别音的实际识别试验，精选特征组合。

（2）合宜采用特征空间中的聚类及向量量化（VQ）技术。

（3）在识别过程中，在DTW、HMM等的基础上，识别规律中增加由先验知识提供的条文（或用基于知识或规律的系统）。

（4）在识别部分合理采用多级识别体制。

（5）设置适当的预处理电路，特征提取后处理部分，如频率提升电路，特殊降噪电路，听觉处理电路或程序等。

（五）人们往往希望减少识别所需要的时间，使在较大字表情况下能够实时完成识别。解决此问题可从下述方面入手：

（1）在软件方面尽量采用快速处理程序。

（2）建立以快速处理片（如TMS320系列，专用乘法器，浮点运算器等）为主体的识别机硬件。

（3）采用并行、锥形等同时处理的多处理器系统。

（4）采用易于大规模集成的统一处理单元，如应用神经网络单元以神经网结构完成识别程序及利用周期阵列器件等。

（5）制造专用硬件（如距离计算硬件，VQ硬件等）与计算设备配合应用。

我们应注意为了制成的设备具有合理的性能价格比及可靠性，并不一定追求采用最快速的器件，而应在设计时根据系统要求作全面考虑。要点是尽量使电路结构等设计合理，充分发挥经合理选用的器件的潜力。

考虑了上述这些问题后，识别系统的结构应在图1.1基本结构的基础上有所补充。补充

后的结构概要框图如图1.5所示。块3中的后处理部分指可能用到的VQ处理，为提高识别率采用的听觉处理部分等。块4中的加工指模板序列的合并(在DTW型系统中)，态模型参数在多输入序列时的参数调整(在HMM型系统中)。块2则包括可能设置的降噪系统和频率提升电路等。

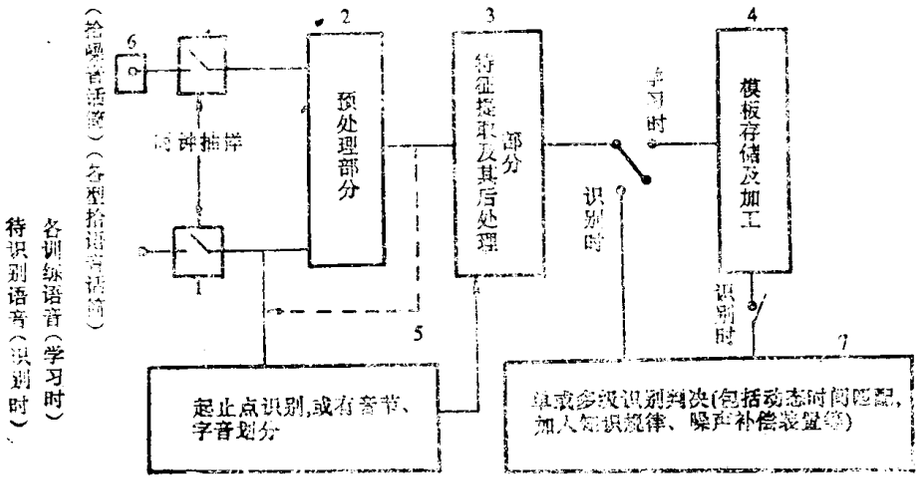


图 1.5

1.4 内容安排

第二章讨论各种有效的特征，尤其是具有新思想和特点的特征，包括其定义，提取方法等。同时也提到各特征宜用的距离定义。内容属于图1.5中的块3、2。所研究的问题包括1.3节中的问题(一)、(四)、(三)。

第三章讨论识别机的各种有效的和具有新思想的学习(训练)方法和识别方法，也包括具有较大实用价值的或新思想特点的起止点判别法及一些“静/虫/清”(S/V/U)短时段划分法。后者在作音节划分或字音划分中有用。适当应用此部分知识(或在加入用语义信息帮助识别的条文)即可将连续音识别问题归化为孤立词音识别问题。本章内容属于图1.5中的块4、7、5，所研究的问题包括1.3节中的问题(二)、(四)等。

第四章讨论多人应用识别系统中为了多人应用增加的处理方法。主要内容属于图1.5中的块4、7，所研究问题包括1.3节中的问题(二)。

第五章讨论在强或较强噪声、干扰环境下识别系统中所采用的处理方法。包括在其中预处理部分等几个块的安排考虑。其内容属图1.5中的块2、7、6、3，所研究的问题包括1.3节中的问题(四)。

第六章讨论给定识别系统具体的要求时，确定宜用的特征及其他系统参数的模拟试验研究方法和所得到的有关选取特征的结论，这往往是具体设计中不可少的部分。其内容主要属于图1.5中的块3等，所研究的问题包括1.3节中的问题(一)、(四)。

第七章讨论语音实时识别系统设计中的总体思想、设备总体结构、分结构、硬件及与之配合的软件的设计问题，并给出一典型系统的实例。在本章中综合应用了上述各章主要骨干部分的内容并作适当补充。主要拟通过它说明作整个系统的设计时应全面考虑的问题和设计的思路及方法。本章所研究的问题主要属1.3节中的问题(五)。

在本书末列有各章有关的参考文献名称，以便进一步查阅。

第二章　特征提取及特征间的距离量度

（一）特征选择的标准：特征的选择对识别效果至关重要。选择的标准应体现对于异字音，相应特征间的距离应大，而对于各同字音，彼此距离应小。若以前者距离与后者距离之比为优化准则用的“目标量”，则应使此量最大。

（二）特征数问题：特征数应尽量取少些以减少计算量，但为了保有高的识别率，所选各特征彼此间的相关性宜小且每个特征的有效性应高（指单用此特征时优化准则中的目标量大）最直接的特征是语音时间波的各脉冲调幅（PAM）值，但相应的特征数量大（如短时段为10ms采样率为10kHz时，每短时段的特征数达100），故应将其转化为较少数的有效者。其思路及注意点有：

1．采用有效且最好有快速算法的一些转换将原语音时间列换为其他各型式的向量（有效的转换指转换后信号能量分布较集中，显出的某些特征较突出），仔细考察它或它的一部分同字音而言的相似性及对异字音而言的差异性，优选留下其中目标量较大者，再通过以后的大量识别试验确定。有必要时可将原语音时间列先加上适当的预处理（如滤波、平均、加权等），再按上面所述进行。

2．研究发音或听觉机理，建立电等效模型。理论上说若模型建立合理，其关键参数应能表征各种语音特色的重要特征，且其数量一般会比PAM型特征的少得多。

3．应适当注意所选特征提取时所需的计算量不要过大，以保证识别系统能作实时识别。

4．注意所选特征（及其有关处理）能与语音过程的非平稳性适应。必要时可采用Wigner-Ville等二阶型特征或用“密距平滑法”处理（指将相邻短时段起始时间之差取小，如0.5ms（此时短时段高度重叠），再将若干个（如8个）相邻短时段算得的特征作平均。以此作为系统的短时段特征存储以压缩所需比较运算的总特征数）。

5．短时段划段的起始时间相对于语音而言常具随机性。应注意对于同一语音时间列，短时段的起始点不同时，所得的特征值差异是否太大。对于不同特征此差异一般不同，所以这点也应为选用特征类型时应考虑的一个因素。下面举一例子。LPC特征属于在激励脉冲影响下的时变滤波器参数的测度。由于在某短时段中发生的激励脉冲个数（如1或2个）及脉冲在短时段中的相对位置与短时段划分起始位置有关，而后者是随机的，所以这会致使同一字音几次测得的LPC特征有较大差异。但是，由于在倒谱特征向量的各元中，表征激励脉冲与时变滤波器参数的元是分离的，所以采用倒谱特征时只用表征滤波参数的元则可使上述的差异减小。

2.1　短时段（帧）能量类特征

（一）短时段帧能量特征（正实数标量）

E_n表征由时刻$n-N+1$到n组成的短时段的能量特征（n为时序号），其定义为

$$E_n \triangleq \sum_{m=-\infty}^{\infty} [x(m)\cdot w(n-m]^2 \qquad (2.1)$$

其中 $x(n)$为离散语音信号时间序列，$w(n)$为时窗函数(其有效长度为N)。E_n也可写为

$$E_n=\sum_{m=-\infty}^{\infty}x^2(m)\cdot h(n-m) \tag{2.2}$$

其中 $h(n)\triangleq w^2(n)$ (2.3)

窗函数可为矩形窗、Hamming 窗等。也可在频域中用卷积法求 E_n，此时将频域信号与低通滤波器的(频域)传输函数相卷积。此低通滤波器可用 FIR 或 IIR 型，不一定相应于矩形或 Hammig 时窗。

(二)对数能量特征 LE_n(实数标量)

$$LE_n\triangleq \lg E_n \tag{2.4}$$

(三)平均幅度函数特征(正实数标量)

$$M_n\triangleq\sum_{m=-\infty}^{\infty}|x(m)|\cdot w(n-m) \tag{2.5}$$

与能量特征比较，其特点为对于大信号的过分夸张现象可以减弱。

(四)平滑信号的能量特征 ES_n

同 E_n，但 E_n 定义式中的 $x(n)$是经过某种规定的低通滤波(即予以平滑)后的信号。

2.2 相对瞬时值类特征

(一)短时段(平均)过零率特征(ZCR 或 NZ)，(正标量)

$$ZCR_n=\sum_{m=-\infty}^{\infty}|\mathrm{sign}[x(m)]-\mathrm{sign}[x(m-1)]|\cdot w(n-m) \tag{2.6}$$

其中
$$\mathrm{sign}[x(n)]\triangleq\begin{cases}1(\text{当 } x(n)\geqslant 0)\\-1(\text{当 } x(n)<0)\end{cases} \tag{2.7}$$

及
$$w(n)=\begin{cases}1/2N(0\leqslant n\leqslant N-1)\\0\ (n\text{ 为其它})\end{cases} \tag{2.8}$$

(二)平滑信号过零率特征 NZS

定义式同上，但其中的 $x(n)$为经规定的平滑处理之后的信号。

(三)帧内最大、最小幅值差特征 ML(正标量)

某帧内各元中的最大值与各元中的最小值之差定义为该帧的现述特征。

(四)一阶差分信号的绝对能量特征 ED

计算出某信号的一阶差分信号，以此作式(2.1)中的 $x(n)$，即得该信号在相应帧中的 ED 特征。

(五)一阶差分信号的最大最小幅值差特征 MLD

(六)一阶差分信号的过零率特征 NZD

(七)平滑信号的帧内最大、最小幅值差特征 MLS

2.3 线性预测系数、预测误差及自相关系数特征[41]

2.3.1 线性预测系数(LPC)特征[41]

(一)方法，概念

LPC为用线性预测法分析语音时得到的有关语音邻样值间某些相关特性的参数组。线性预测分析基于如下的基本概念，即一语音样值能用过去的若干语音样值的线性组合来近似估计(预测)。按在一所分析的帧(短时段)内实际的各语音样与各预测得的样之间差值的平方和最小准则，可以决定唯一的一组预测系数，即LPC。

设$\{x(n)\}$为语音时间序列，第n个语音样可用前p个语音样来预测，即$x(n)$的预测值

$$\hat{x}(n) = -\sum_{i=1}^{p} a_i \cdot x(n-i) \tag{2.9}$$

其中$a_i(i=1\cdots p)$为预测系数，而预测误差为

$$e(n) = x(n) - \hat{x}(n) = \sum_{i=0}^{p} a_i \cdot x(n-i) \tag{2.10}$$

其中$a_0=1$为固定常数。

称一帧内各时刻误差的平方和为E，则对于具有N个样点的帧而言

$$E = \sum_{n=0}^{N-p-1} \left[x(n) + \sum_{i=1}^{p} x(n-i) \right]^2 \tag{2.11}$$

前已谈到决定a_i的准则是使E极小，故令$\partial E/\partial a_i = 0(i=1\cdots p)$，于是可得到一线性方程组：

$$\sum_{j=1}^{p} a_i \cdot R(i-j) = -R(j) \qquad (j=1\cdots p) \tag{2.12}$$

其中

$$R(j) \triangleq \sum_{n=0}^{N-1-j} x(n) \cdot x(n+j) \tag{2.13}$$

称为语音短时自相关系数。N为语音帧中样点数，p为预测器阶数。解方程式(2.12)可得$a_i(i=1\cdots p)$，其各解法叙述如下：

1. Durbin算法：式(2.12)为Toeplitz型线性方程组，故可用Durbin解法[41]。它是一种“阶”递归算法，其步骤为：

(1) 初始化：令$I_0 = R(0)$

$$k_1 = R(1)/R(0) \tag{2.14}$$

$$\alpha_1^{(1)} = k_1 \tag{2.15}$$

(2) 第一循环运算(称为A循环)：以$i=2\cdots p$，按下面各式作循环运算：

$$I(i-1) = \left(1 - k_{i-1}^2\right) \cdot I(i-2) \tag{2.16}$$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} \cdot R(i-j) \right] / I(i-1) \tag{2.17}$$

$$\alpha_i^{(i)} = k_i \tag{2.18}$$

（3）第二循环运算(称为 B 循环)：以 $j=1\cdots i-1$，按下式循环运算：

$$\alpha_j^{(i)}=\alpha_j^{(i-1)}-k_i\,\alpha_{i-j}^{(i-1)} \tag{2.19}$$

（4）最后，令

$$a_j=-\alpha_j^{(p)} \qquad (1\leqslant j\leqslant p) \tag{2.20}$$

a_j 即待求的各 LPC 系数。

2．格型算法：此算法形式上与上述算法有较大差异，它是先设想令信号行经一种格型结构的预测误差滤波器，按使格型结构各级出端信号(误差信号)在短时段内均方和最小的准则，算出每级交叉支路中的乘子 $k_1\cdots k_p$（p 为级数或阶数，k_i 等称为偏自相关系数或反射系数），这也就使各级出端信号的自相关性或可预测性逐级减小。因为 k_i、a_i 分别是在格型与横向型结构中对输入信号作预测时所用的参数，故对于同一输入信号，其 k_i 与 a_i 参数会有一定联系利用导得的联系式即可从 k_i 算出所需的 a_i。上述的格型结构如图2.1所示。

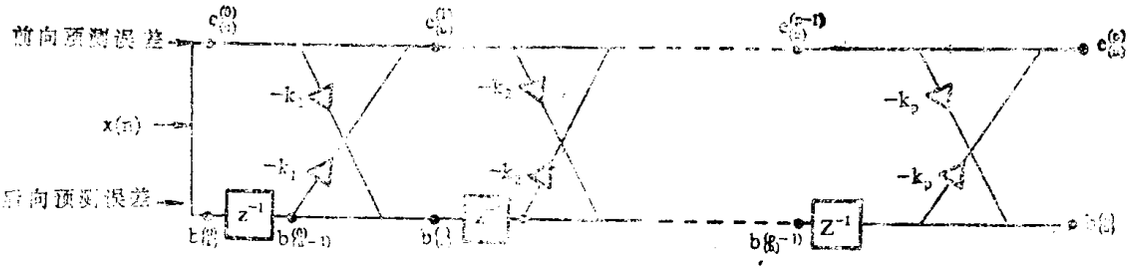


图 2.1

由图2.1可知

$$e^{(i)}(n)=e^{(i-1)}(n)-k_i\cdot b^{(i-1)}(n-1) \tag{2.21}$$

$$b^{(i)}(n)=b^{(i-1)}(n-1)-k_i e^{(i-1)}(n-1) \tag{2.22}$$

$$e^{(0)}(n)=b^{(0)}(n)=x(n) \tag{2.23}$$

由使前向预测误差平方和最小为准则，从上述关系可导得，满足准则的

$$k_i=\sum_{n=0}^{N-1}e^{(i-1)}(n)\cdot b^{(i-1)}(n)\left[\sum_{n=0}^{N-1}e^{(i-1)}(n)^2\cdot\sum_{n=0}^{N-1}b^{(i-1)}(n)^2\right]^{1/2} \qquad (i=1\cdots p) \tag{2.24}$$

Burg 提出另一准则，即是使图2.1中前向与后向预测误差的平方和最小。也即令

$$\tilde{E}^{(i)}\triangleq\sum_{n=0}^{N-1}\left[e^{(i)}(n)^2+b^{(i)}(n)^2\right] \tag{2.25}$$

最小。此时相应的

$$k_i=2\sum_{n=0}^{N-1}\left[e^{(i-1)}(n)\cdot b^{(i-1)}(n)\right]\left[\sum_{n=0}^{N-1}[e^{(i-1)}(n)]^2+\sum_{n=0}^{N-1}b^{(i-1)}(n)\right]^2 \qquad (i=1\cdots p) \tag{2.26}$$

总结上述可得，按格型结构计算 k_i 和各 LPC(a_i)的步骤为：

（1）初始值设为 $e^{(0)}(n)=b^{(0)}(n)=x(n)$；

（2）按式(2.26)计算 k_1，并令它等于 $\alpha_1^{(1)}$；

（3）按式(2.21)、(2.22)算出 $e^{(1)}(n)$、$b^{(1)}(n)$；

（4）设 $i=2$；

(5) 由式(2.26)，得到 $k_i=\alpha_i^{(i)}$；

(6) 由前面 Durbin 算法中用过的式(2.19)算出 $\alpha_j^{(i)}(j=1, \cdots i-1)$，它们即线性预测系数 $aj(j=1, \cdots i-1)$；

(7) 由式(2.21)、(2.22)算出 $e^{(i)}(n)$，$b^{(i)}(n)$；

(8) 设 $i=i+1$；

(9) 若 $i\leqslant p$，返回第 5 步骤；

(10) 结束。

本法的特点是可以从语音较直接地得到 LPC，不需经计算自相关系数此中间步骤。本法可视为方程式(2.12)的一种间接解法。

如果将前述的求 a_i 的准则略加变动，即将式(2.11)中的 E 改为

$$E'\triangleq\sum_{n=0}^{N-1}e^2(n) \tag{2.27}$$

可以导出相应于式(2.12)的含 a_i 的方程组为

$$\sum_{k=1}^{p}a_k\phi_n(i,\ k)=\phi(i,\ 0)\quad(i=1,\ \cdots p) \tag{2.28}$$

其中

$$\phi_n(i,\ k)\triangleq\sum_{m=-k}^{N-k-1}x_n(m)\cdot x_n(m+k-i)\qquad\begin{bmatrix}1\leqslant i\leqslant p\\0\leqslant k\leqslant p\end{bmatrix} \tag{2.29}$$

式(2.28)虽与式(2.12)相差不大，但它却不再是 Toeplitz 方程(对称方程)，故需用其它解法。下面叙述解此种方程的“协方差算法”。

3. 协方差算法：下面讨论用 Cholesky 分解技术的求解式(2.28)的协方差算法[41]。其思路如下：

令 $\alpha\triangleq(a_1\cdots a_p)^T$，将式(2.28)写为矩阵形式

$$C\alpha=\Psi \tag{2.30}$$

考察得知 C 必为一对称正定阵，故可用“Cholesky 分解法”(或称“方根法”)将 C 分解成

$$C=VDV^T \tag{2.31}$$

形式，其中 V 为对角线元均为 1 的下三角阵，D 为对角阵。将上式代入式(2.30)得 $VDV^T\alpha=\Psi$。此式可分裂为二

$$\begin{cases}VY=\Psi & (2.32a)\\ DV^T\alpha=Y,\ \text{即}\ V^T\alpha=D^{-1}Y & (2.32b)\end{cases}$$

Y 为引入的参变量。这样一来原题即被分成了两个简单的题。对于第一题，由于 V 为三角阵故可将 Y 向量的各元递推解出。于是式(2.32b)左侧成为已知。按同样思路由式(2.32b)又可将 α 的各元，所需的 $a_1\cdots a_p$ 解出。

至于由 $C=[c_{ij}]$ 得到 $V=[v_{ij}]$，$D=[d_i]$ 中各元的方法，因 V、D 形式简单，故也易用将式(2.30)展开、通过对照递推求解的方法。举例说，对于 $p=4$ 情况，此时 $i=1$，2，3，4，经如上述对照后求解，可得

$$\begin{cases} d_1=\phi_{11},\ d_2=\phi_{22}-V_{21}^2 d_1 \\ V_{21}=\phi_{21}/d_1,\ V_{31}=\phi_{31}/d_1,\ V_{41}=\phi_{41}/d_1 \\ V_{32}=(\phi_{32}-V_{31}d_1V_{21})/d_2,\ V_{42}=(\phi_{42}-V_{41}d_1V_{21})/d_2 \end{cases} \tag{2.33}$$

由式(2.32a, b)递推解出 $a_i(i=1\cdots4)$的递推式则为

$$\begin{cases} Y\Psi_{11} \\ Y_i=\Psi_i-\sum_{j=1}^{i-1}V_{ij}Y_j \quad (4\geqslant i\geqslant 2) \end{cases} \tag{2.34}$$

而

$$\begin{cases} a_4=Y_4/d_4 \\ a_i=Y_4/d_i-\sum_{j=j+1}^{4}V_{ji}a_j \quad (1\leqslant i\leqslant 3) \end{cases} \tag{2.35}$$

用式(2.11)与用式(2.27)作为建立准则的差别是，对于后者得到的数据一般较前者准确，但缺点是结果不能保证稳定，只有当每帧采样数足够大时，才能保证满足实际需要的稳定性。

(二)数据加窗问题

在语音帧的起止点，线性预测方法以零值预测非零值或以非零值预测零值一般会使预测误差增大。若对数据先作适当的加窗处理，可减小上述预测误差。即是说(当选用 Hamming 窗时)先用

$$w(n)\triangleq 0.54-0.46\cdot\cos[2\pi n/(N-1)] \quad (n=0,\ 1,\ \cdots N-1) \tag{2.36}$$

乘 $x(n)$，得到加窗后的数据

$$S(n)=x(n)\cdot w(n) \quad (0\leqslant n\leqslant N-1)$$

(N 为帧内数据个数)再以 $S(n)$代替 $x(n)$用上述各法进行 LPC 特征计算。

(三)LPC 特征的距离量度

合理拟定距离定义应掌握的一个主要原则是：设令某 LPC 组(特征)以其各元为线性预测的权值预测某信号 $x(n)$，按式(2.11)得一预测误差平方和 E_1；又令另一 LPC 组按同法预测同一 $x(n)$得到一预测误差平方和 E_2。则若 $E_1=E_2$，按拟定的距离定义求得的距离应为 0；否则所拟定义不合理。下面叙述较实用的几种距离定义。

1. Itakura 增益归一化距离

(1) 定义：设一帧信号 $\{x(n)\}$，其自相关向量 $R=[R(0),\ R(1),\ \cdots R(p)]^T$。信号经 LPC 分析后，得到预测系数向量 $A=[a_0,\ a_1,\ \cdots a_p]^T$，显然 A 是平方误差和最小意义的最佳预测系数(特征)。若以某系数 $A'=[a_0',\ a_1',\ \cdots a_p']^T$ 代替 A 对$\{x(n)\}$作预测，则将使预测的平方误差和增大。现定义 A 与 A' 间的“Itakura 增益归一化距离”

$$d(A,\ A')\triangleq\frac{\alpha}{\alpha_M}-1 \tag{2.37}$$

其中

$$\alpha\triangleq R(0)R_a(0)+2\sum_{j=1}^{p}R(j)R_a(j) \tag{2.38}$$

$$\alpha_M\triangleq R(0)+\sum_{j=1}^{p}a_j\cdot R(j) \tag{2.39}$$

而

$$R_a(i)\triangleq\sum_{j=0}^{m-i}a'_j\cdot a'_{i+j} \tag{2.40}$$