分类号　　密级：

清華大學

硕 士 学 位 论 文

题　　目 非特定人连续数字识别方法与汉语语音数据库的研究

并列英文题目 **Studies on Speaker-Independent Continuous Digit Recognition Methods and Chinese Speech Corpus**

研究生姓名　郑　方　　**系别**　计算机科学与技术系

专　　业　计　算　机　应　用

导 师 姓 名　吴文虎　　**职称**　教　授

论 文 答 辩 日 期　一九九二年六月

摘 要

本文对非特定人连续数字识别方法进行了研究。作者在经过大量实验之后，确定了一种识别算法：首先，使用基于非线性分块的分段概率模型，在识别时，参考模型各状态的分数估算用段长度进行规整(平均)；其次，使用声调(概率)评价函数，这个函数给出了某数字为三类声调(一声类、二三声类和四声类)中每一类的概率；同时，考虑到连续语音中毗邻音节之间相互影响，采用与音节所处位置有关的多套模型。使用上述的策略建立的系统，其识别效果有相当大的改善。

本文还介绍了一个以汉语普通话为基础的汉语语音数据库，这样的数据库，对语音识别、语音分析甚至语言理解方面的研究工作都将有很大帮助。

关键词：非特定人连续数字识别，非线性分块，分段概率模型，声调(概率)评价函数，汉语语音数据库

ABSTRACT

In this paper, the methods of Speaker-Independent Continuous Digit Recognition are studied. After a number of experiments, a kind of recognition algorithm is proposed. Firstly, it adopts the Segmental Probability Model based on the Non-Linear Partition Principle, when the syllable to be recognized is matching to each reference model, the score estimation of each state is normalized (averaged) by the length of the corresponding state observation sequence; secondly, it adopts the tone (probability) discrimination function, which gives the syllable to be recognized the probability values of three corresponding tone classes (the 1st tone class, the 2nd and 3rd tone class, and the 4th tone class); in the meantime, in consideration of the interaction between the adjacent syllables in the continuous speech, it adopts syllable-position-dependent Multi-Set-Models. All these above strategies have improved the recognition performance of the established system greatly.

On the other hand, a Chinese Speech Corpus based on the Chinese Mandarin is introduced in this paper, such a corpus will be of a great help to the research in fields of speech recognition, speech analysis, and even lanuage understanding.

Fang Zheng (Computer Application)
Directed by Professor Wenhu Wu

Keywords:
Speaker-Independent Continuous Digit Recognition, Non-Linear Partition, Segmental Probability Model, Tone (Probability) Discrimination Function, Chinese Speech Corpus

目录

第一章　综述

语言是人类获得信息的主要来源之一，是人与外界交流信息的最方便、最有效、最自然的工具。随着计算机科学与应用的发展，出现了计算机语音学(Computer Phonetics)。人们对计算机语音的研究主要有以下几个方面：

- 语音编码 (Speech Coding)；
- 语音合成 (Speech Synthsis)；
- 语音识别 (Speech Recognition)；
- 话者识别 (Speaker Recognition)或
  话者确认 (Speaker Verification)。

语音识别(SR)就是让计算机听懂人说话，它是发展人机声通信和新一代智能计算机的重要组成部分。它有几种分类方法：按被识别人的范围可分为特定人(Speaker Dependent)和非特定人(Speaker Independent)语音识别；按词汇量的大小可分为小词汇量(Small Vocabulary)和大词汇量(Large Vocabulary)语音识别；按说话方式可分为孤立词(Isolated Word)和连续语音或连接词(Continuous Speech or Connected Word)语音识别。它们的难易程度如表 1.1 所示(E：易， D：难)。

表 1.1　语音识别难易评价表

适应对象	词汇量	识别方式	难易评价	说明
特定人	小词汇量	孤立词	EEE	较易
		连续语音	EED	较难
	大词汇量	孤立词	EDE	较难
		连续语音	EDD	很难
非特定人	小词汇量	孤立词	DEE	较难
		连续语音	DED	很难
	大词汇量	孤立词	DDE	很难
		连续语音	DDD	极难

§1.1　SR 的早期研究(76 年以前)

早在 60 年代末期，面对语音识别的重重困难，人们试图对语音识别的任务作一简化，即不急于识别由任何人、以任何方式说的任何内容的连续语音，而是首先解决一个子问题：特定人、小字表、孤立词，从而使 SR 研究能在当时的技术条件下得以开展。这在 70 年代中期取得了长足的进展：

1．在语音信号表示和特征抽取方面提出两种表示法：

以滤波器组输出或 FFT 系数这些领域特征作为特征参数。

以线性预测编码(Linear Predictive Coding)分析为基础的特征参数：LPC 参数、CEP 系数(倒谱 Cepstrum 系数)、部分相关系数、声道面

积、Mel-CEP 系数等等，以及相应的相似度测量。

2. 以动态规划(Dynamic Programming)为基础的模板匹配技术的出现[Vintsjuk 68]，使得在此后近十年内，人们一直视动态时间弯折(Dynamic Time Warping)为主要方法，并使 SR 实用化成为可能。

3．以人工智能(Artificial Intelligence)为基础的 DARPA(Defense Advanced Research Prejects Agency)语音理解计划[Klatt 80]，把高层知识用于 SR，但在实时性、实用性和鲁棒性方面不理想。

§1.2 SR 的中期研究(77 年~82 年)

七十年代后期，当特定人、小词汇、孤立词 SR 达到令人满意的结果之后，人们开始沿三个不同方向拓展研究领域和目标：

1．特定人向非特定人拓展

采用 K-mesns 聚类算法对多个人的发音样本进行聚类。

2．孤立词向连接词拓展

提出了 Level-Building [Myers 81], Two-Stage DP[Sakoe 79], One-Pass DP[Bride 82]等基于 DP 的新的匹配方法。

以上两个扩展基本上都是基于小词汇量，尤其是数字识别(0-9)。

3．小词汇量向大词汇量拓展

这一扩展遇到了计算量和存储量急剧增加的困难，相应出现了以下方法：

矢量量化(Vetcor Quantization)技术[Linde 80]：它具有很好的数据压缩能力及理想的聚类功能，因此人们将 VQ 用于 SR 进行预处理或预选，以减少识别运算量。

子词单元(Subword，如音节、音素等)的提出和应用：主要用以减少运算量和存储量。

采用**分级识别(粗分类)**进行预选[Rabiner 81].

§1.3 SR 的近期研究(83～89 年)

80 年代中期以来，新技术的不断出现使语音识别有了实质性的进展，特别是隐马尔可夫模型(HMM)的广泛研究和应用，使语音识别能同时在大词表、非特定人、连续语音三个方面取得重要发展。

1．HMM

最早将隐马尔可夫模型用于 SR 是 70 年代中期[Baker 75, Jelinek 76]，但对 HMM 的全面研究和大规模应用是 80 年代以后的事。它受到广泛重视的原因是：

马尔可夫链可以用来描述蕴藏于观察数据中的时变特性，这使得它能处理语音信号中常常出现的非平稳特性(即时变特性)。

它不仅能用于描述各种不同层次的语音单元，甚至可以描述 VQ 中的一个码字或由声学特征定义的任一种声学单元，并且由小单元模型组成大

单元模型(音节(或音素)→单词→句子)。

由 Viterbi 解码可得到与语音序列对应的最佳状态序列，从而得到语音单元的最佳分割，使子词单元的使用非常方便，大大避免了训练和识别时的分割困难，使连续语音识别问题得到解决。

随着对 HMM 的深入研究和在 SR 中的需要，许多新的算法产生，如 MLE 估计、平滑、外插、建立时间模型、话者自适应等等，使得这一技术在 SR 中有了更深入的应用。

2．神经网络(Neural Networks)

80 年代中期重新开始的 NN 研究，也给 SR 带来一片新的生机。由于 NN 具有自组织和自动学习各种复杂分类边界的能力，以及很强的区分能力，使它特别适用于 SR 这一特殊的分类问题。人们将 NN 和 HMM 在同一 SR 系统中结合使用，即由 NN 完成静态的模式分类问题，而用 HMM 甚至传统的 DP 来完成时间对准问题[Sakoe 89, Gao 90, Morgan 90, Franzini 90]。从实验结果来看，这种思想可行而且有效，并能使 NN 比较容易地用于连续语音识别问题[Morgan 90]。

语音识别常用的 NN 有：

时间延迟神经网络 TDNN [Waibel 88]
递归神经网络 RNN[Bourlard 89]
连接预测神经网络 LPNN[Tebelskis 90]
自组织神经网络 SONN[Kohonen 84]
学习矢量量化 LVQ[McDermott 89]
混合语音识别系统

3．基于知识的 SR

与上述基于统计分析和强有力的算法的研究几乎并行开展的是以 MIT 的 V.M.Zue 教授[Zue 85]、McGrill 大学的 De Mori 教授[De Mori 86]和法国 CRIN/INRIA 的 J.P. Haton 教授[Haton 84]为代表的基于语音学知识的 SR 研究。MIT 的 SUMMIT 系统[Zue 89, Zue 90]则是基于知识的 SR 的典范，它实现了非特定人、大词汇量和连续语音的识别。

§1.4 数据库和复杂性度量

与识别方法的研究同样重要的是必须建立标准的实验室数据库，以便进行各种方法之间、不同识别系统之间的比较。目前已有的非特定人、大词汇量、连续 SR 系统几乎全部采用 DARPA 的含有 997 个词的海军资源管理数据库。另一个非常著名的数据库是 TIMIT，由于 TI 公司与 MIT 公司联合研制。

一个识别系统的优劣评定也是一个重要问题。Jelinek 基于信息论提的 Perplexity 是描述任务复杂性的一种好的度量[Bahl 80],它可描述复杂性与识别率之间的关系。在实际描述一个词汇表时，Perplexity 甚至比词汇量更重要。

§1.5 国内 SR 研究现状

国内 SR 研究起步较晚，除中科院声学所外，大多数单位是 70 年代末及 80 年代初才开始。但在短短的十年间，已取得重大进展，在孤立词(多音节)识别方面，清华大学、中科院声学所及哈工大完成了较高水平的识别系统。近年来，国内一些单位致力于研究和建立以全音节识别为基础的大词汇识别系统，中科院自动化所、四达公司、星河公司都建立了上万词的识别系统，并向实用化迈了可喜的一步。清华大学的3400成语非特定人识别系统，它使用多码本矢量量化(MCVQ)技术进行汉语全音节识别，也有令人鼓舞的效果[Li 90]。

与国外工作相比，国内另一个不足之处是没有公认的标准语音数据库，这也是一个需要解决的问题。

第二章语音识别的基本知识

§2.1 数据的采集及前端处理

§2.1.1 数据采集

把语音从模拟量(信号)转换成数字量(信号)并存储到计算机中去，这就是语音采集的任务，这任务由 A/D 转换电路和语音采集程序完成。进行语音采集之前首先应该确定采样频率，采样频率的确定必须满足采样定理：采样频率必须是语音所含的频带的两倍以上。一般地，人类的语言信号的频率大多集中在 50～5KHz 范围内，因此，采样频率应在 10KHz 左右。

确定了采样频率，那么必须去掉语音信号中频率为采样频率的 1/2 以上的高次谐波成分。去掉高次谐波的滤波器称为去伪滤波器(Anti-alias Filter)，这就是为进行高精度数字语音分析或合成所必须的滤波器。

TEXAS 生产的 TMS320 系列的快速信号处理板(FSP)上有一个截止频率为 4.5KHz 的去伪滤波器，采样频率由软件设定为 9.6KHz 等。TMS320C25 的内部中断 XINT 或 RINT 为语音的采集和存储提供了精确的定时和方便的存取[TI 88，TMS 89]。

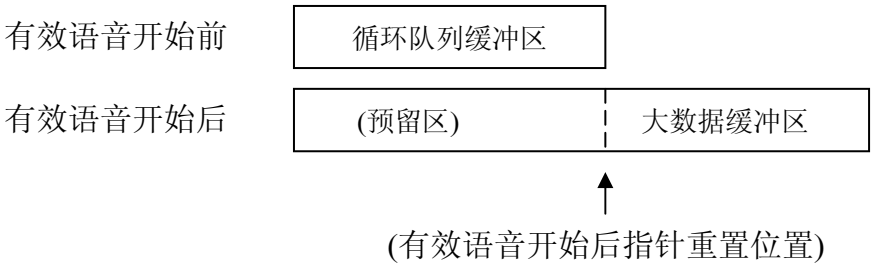

图 2.1 语音循环队列缓冲区

为了充分保留某些辅音和过渡音的信息，以及进行有效语音的开始点的判定，我们采用了循环队列技术[Zheng 90]，如图 2.1，有效语音开始前的缓冲区为一循环队列，它的长度根据需要可定为 10～20 帧，当有效语音开始后，信息被存在一个比较大的缓冲区中去。主程序负责把循环队列中的信息正确地搬到这个大的缓冲区头部的预留区，并给出数据的首地址。

§2.1.2 短时帧平均幅度及短时帧过零率的统计

语音是一种复杂的时变信号，但我们可以认为在极短的时间内，语音信号是线性时不变系统的输出，这就是语音的短时分析方法。即把语音信号在窗函数的作用下，一段一段地进行处理。帧是进行短时分析的语音长度单位，一般为窗的宽度，相邻帧之间一般有交叉，两帧之间起始点的距离称为帧移。一般地，取帧长为 30ms 左右，帧移 10ms 左右。在我们的系统中，帧长为 256 点，帧移为 128 点。

在短时语音分析中，帧平均幅度及帧过零率都是很重要的指标量，它们是这样进行统计的：

- **帧平均幅度**是指一帧语音的平均幅度，计算公式为

$$\overline{A} = \frac{1}{N}\sum_{i=0}^{N-1}\left|S_n(i)\right|$$

- **帧过零率**是指一帧语音的短时过零数，计算公式为：

$$FZRO = \sum_{i=0}^{N-1}\xi_i$$

其中，

$$\xi_i = \begin{cases} 1, & \text{当}\left|S_n(i)\right| \geq ZLEV\text{而且}S_n(i) * S_n(i-1) < 0\text{时} \\ 0, & \text{其它} \end{cases}$$

公式中，N 为帧长(窗宽)；$S_n(i)$ 表示以 n 时刻为起点的 i 时刻的数字化语音样值；$ZLEV$ 是用于统计过零率的阈值(本文称作**零电平宽度**)，使用它可减少噪声对帧过零率的影响，而且过零率在音节的切分点处及无声段处比其它地方要低。这是一个经验值，可以通过噪音统计得出。

§2.1.3 端点检测及音节切分

所谓语音的端点检测，就是语音的首尾判定，它是把一段语音定为有效语音段的粗判，是进一步进行有效语音段细判和字词分割的基础。语音的端点检测与语音采集同时进行。

作为进行语音的端点检测的指标量，有好几种可供选择。比较常见的有利用帧平均幅度或帧过零率来进行判定的，也有利用两者综合判定的。选择指标量的原则是：一要尽量准确，二是简便易行。

在我们的系统中，我们根据帧平均幅度进行端点检测，而且帧平均幅度、帧过零率及音长信息为判据，使用相对阈值进行音节切分[Zheng 92]，收到很好的效果。

§2.1.4 低通滤波

原始语音一般要经过一个低通滤波器用以去掉不必要的高频噪音，我们

的系统使用[Zhong 88]介绍的简单整系数低通数字滤波器进行滤波。

这个滤波器的转移函数为：

$$H(z)=\left(\frac{1-Z^{-M}}{1-Z^{-1}}\right)^{K}$$

其中，M 是单位圆周上的零点数(不计多重零点)，K 为数字滤波器(Digital Filter)的阶数。K 的引入可以使通带下降沿变得更加陡峭，低通性能更好。

这样的 DF 需要确定下列参数：

ω_p(孤度)：通带截止频率

α_p(dB)：通带内最大衰减

α_s(dB)：阻带最小衰减

f_s(Hz)：采样频率(对应角频率为 $\omega_s=2\pi$)

其中 α_p 定义为：

$$\alpha_p=201g\left|\frac{H(e^{j0})}{H(e^{j\omega_p})}\right|=201g\frac{1}{\beta}$$

习惯上它取 3dB；α_s 定义为：

$$\alpha_s=201g\left|\frac{H(e^{j0})}{h_1}\right|$$

这里 h_1 为频响第一旁瓣的峰值。

在这些参数确定后，DF 的幅频特性就确定了，图 2.2。

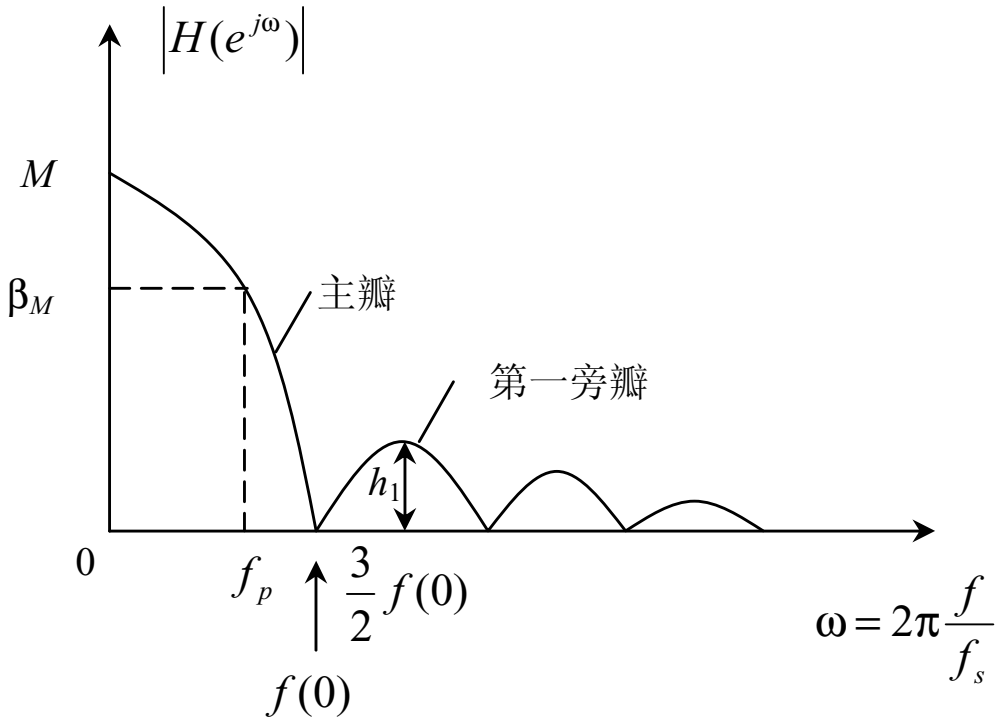

图 2.2　低通数字滤波器幅频特性

[Zong　88]给出了 K、M 的近似计算公式：

$$K = 0.07427\,\alpha_s$$
$$M = f_s / \left(2.25 f_p \sqrt{K}\right)$$

§2.1.5 原始语音的预加重及加窗处理

为了进行高频提升，在对原始语音进行处理之前，先进行预加重(Pre-weighting)处理。预加重处理的公式是：

$$S_n'(i) = S_n(i+1) - 0.95 S_n(i)$$

这相当于在原始语音进行处理之前，先让它经过一个滤波器，该滤波器的系统函数为：

$$H[z] = 1 - 0.95 z^{-1}$$

对建立在短时分析基础上的各种语音处理，加窗是必要而且重要的。在窗的范围内我们认为语音是相对稳定的，也即语音参数是稳定的，这是短时语音分析的基础。窗有几种形式，最常见的有**矩形窗**：

$$W_r(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & \text{其它} \end{cases}$$

和**哈明窗**：

$$W_h(n) = \begin{cases} 0.54 - 0.46\cos[2n\pi / N], & 0 \le n \le N-1 \\ 0, & \text{其它} \end{cases}$$

[Rabiner 78]认为哈明窗有更好的频域特性。

§2.2 特征抽取及特征空间的距离质量

特征参数的抽取是对语音信号进行有效压缩和进行语音识别的特别重要的一环。对于任一段特定的语音，特征参数必须含有该语音的信息并具有相对的稳定性。在语音识别的特征抽取中，大致可分为两类。五六十年代，当语音识别学科刚刚起步的时候，基于信号处理学科所取得的成就，以频谱(Spectrum)分析为基础的语音特征得到了普遍的应用；六十年代后期以来，线性预测编码(Linear Predictive Coding)分析方法被引入语音处理领域，从而以 LPC 分析为基础的特征占主导地位。总的说来，对后一类特征参数的评价较高，尤其是基于同态分析的倒谱(Cepstrum)系数，它不仅在特定人语音识别中取得了很好的效果，而且在非特定人的语音识别系统及说话人的识别中也被公认为最有效的特征表示。

§2.2.1 线性预测的基本原理

“线性预测”(Linear Prediction)这个术语最早由 N. Wiener [Wiener 66]提出，此后这一技术在很多领域起到了很重要的作用。它首先由 Itakura 和 Saito [Itakura 68]及 Atal 和 Schroeder [Atal 68]引入语音识别与合成领域，并对语音研究的各个

方面产生了极大的影响[Mardel 76]。它的基本假设是：语音的波形和频域特征可以由一组很少的参数来有效而精确地表征，而这些参数只需经过简单的计算即可获得。从这里可以对线性预测的重要性略见一斑。

语音抽样不仅与前一时刻的样值有关，而且与前几个时刻的语音样值都有关。线性预测模型假定，现时刻的语音样值可以用前 p 个时刻的样值的线性组合表示，即：

$$\hat{S}_n(m)=\sum_{k=1}^{p}a_k S_n(m-k)$$

这就是线性预测的基本原理。其中，$S_n(m)$ 表示为 n 时刻为起点的 m 时刻的数字化语音样值，a_1，…，a_p 称为线性预测(LPC)系数。

如果把语音的发声机制看作一个数字滤波器的话，很容易证明，使用 LPC 系数可以推出这个时变滤波器的极点[Rabiner 78]。

$$H(z)=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}}$$

在这个模型中，浊音是由周期性脉冲信号激励产生，清音则由随机噪音激励产生。

实际语音样值与由线性预测得到的语音预测值的差，称为语音残差：

$$e_n(m)=s_n(m)-\hat{s}_n(m)$$

并定义短时平均预测误差 E_n：

$$E_n=\sum_m e_n^2(m)$$

根据使 E_n 最小的原则，可以使语音的预测值达到与原始语音样值的最佳逼近，从而可以求得一组线性预测系数 $\{a_k\}$。计算预测器系数的方法很多，常见的算法有自关法、格型法和协方差法[Rabiner 78]。

本系统采用自关法，它假定波形 $S_n(m)$ 在 0≤m≤N-1 范围内不为 0(N 为窗宽)。显然，对于一个 p 阶的线性预测器，其相应的预测误差只在 0≤m≤N-1+p 范围内不为 0。故有：

$$E_n=\sum_{m=0}^{N-1+p}e_n^2(m)$$

分别令该式对 a_k 的偏导数为零，可得到一个方程组：

$$\sum_{k=1}^{p}a_k R_n(|i-k|)=R_n(i),\qquad 1\le i\le p$$

该方程组可用矩阵的形式表示，称为自相关阵。这个自相关阵是一个托伯利兹矩阵，其中 $R_n(i)$ 为自相关系数，定义为：

$$R_n(i)=\sum_{m=0}^{N-1+p}s_n(m)s_n(m+i),\qquad 1\le i\le p$$

[Rabiner 78]介绍了求解该矩阵的杜宾(Durbin)递推法，递推公式为：

$$E^{(0)}=R(0)$$

以下对 i=1，2，…，p 进行递推

$$K_i = \left[R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j) \right] / E^{(i-1)}$$

$$a_i^{(i)} = K_i$$

$$a_j^{(i)} = a_j^{(i-1)} - K_i a_{i-j}^{(i-1)}, \quad 1 \le j \le i-1$$

$$E^{(i)} = (1 - K_i^2) E^{(i-1)}$$

最终解为：

$$a_j = a_j^{(p)}, \quad 1 \le j \le p$$

§2.2.2 倒频谱分析基本原理及距离度量

由语音的产生模型可知，在短时间内，可把语音信号看作是线性时不变系统的输出，它是激励与冲激响应的卷积。

要研究系统的性质，就需把激励与输出分开，而着重分析与系统本身性质有关的冲激响应，这就需要采用一种称为“解卷”或同态分析的方法。倒频谱就由此产生。

一个系统 $H[\cdot]$称为卷积同态系统，是说它具有下式的性质(*为卷积运算符号)：

$$H[x_1(n) * x_2(n)] = H[x_1(n)] * H[x_2(n)]$$

这类系统的特性是某一种分量(所需要的分量)可以基本不变地通过它，而不需要的分量可以被除掉。同态系统理论的一个重要方面是任何同态系统都可以表示为三个同态系统的级联($Z[\cdot] \to \log[\cdot] \to Z^{-1}[\cdot]$) [Rabiner 78]。

所谓同态分析，就是把呈卷积关系的两信号变换为呈线性相加的两信号，再用不同通带的滤波器滤掉不必要的成分。为了便于计算，我们选取在 Z 域的单位圆上进行计算。

根据倒频谱的定义可知，该系统的冲激响应的倒频谱的 Z 变换为 $\log H(z)$：

$$\log H(z) = \sum_{k=-\infty}^{\infty} C_k z^{-k}$$

这里$\{C_k\}$称为倒频谱系数。求倒频谱的计算量很大，但可以证明，由 LPC 系数可以推出倒频谱参数(CEP 系数)。因稳定的系统的极点均在单位圆内，此系统没有零点，因此为最小相位系统。$\log H(z)$ 在包含单位圆在内的区域上是解析的。

$$[\log H(z)]' = \frac{H'(z)}{H(z)}$$

将 $\log H(z)$ 及 $H(z)$ 的表达式分别代入上式左右边，并用 LPC 系数近似时变滤波器 $H(z)$ 的参数即可得：

$$C_k = a_k + \sum_{n=1}^{k-1} \frac{n}{k} C_n a_{k-n}, \quad 1 \le k \le p$$

用外推法可近似求得大于 p 阶的倒频谱系数，

$$C_k = \sum_{n=1}^{p} \frac{k-n}{k} a_n C_{k-n}, \quad p \le k \le p'$$

这样我们就得到了 p'阶倒频谱(CEP)系数，它们构成一个 CEP 矢量。在我们的研究中，CEP 系数是主要的特征参数。我们取 p=12, p'=16。由 LPC 系数推导出来的 CEP 系数也叫 LPC-CEP。以后记 $ND = p'$。

关于 CEP 矢量间的距离度量，有以下几种方法，设 X、Y 分别为两个 CEP 矢量：

• 绝对值距离度量

$$d(X,Y) = |X - Y| = \sum_{i=1}^{ND} |X_i - Y_i|$$

• 欧氏距离度量

$$d(X,Y) = \|X - Y\| = \left[\sum_{i=1}^{ND} (X_i - Y_i)^2 \right]^{1/2}$$

• 加权欧氏距离度量

$$d(X,Y) = \left[\sum_{i=1}^{ND} [W_i (X_i - Y_i)]^2 \right]^{1/2}$$

其中 $W = (W_1, \cdots, W_{ND})$ 为权矢量。由于实际语音的 CEP 矢量的各维的幅度大小并不协调，为了使各维对距离度量都有大致相当的贡献，我们选用加权欧氏距离度量。权向量是对大量的 CEP 矢量的各维进行幅度均值统计后确定的。

从人类听觉系统研究的成果来看，可以认为人耳分辨声音频率有一种类似于取对数的功能，或称为 Bark 因子[Zwicker 61]，基于此而出现的 Mel-Cepstrum [Davis 80]比 LPC-Cepstrum 具有更优越的性能；另一种尝试是将双线性变换应用于倒频谱系数[Lee 88a]，也取得了较好的效果。

但 LPC 分析忽视了能量信息，为了弥补这一缺陷，归一化能量常常与 CEP 系数一起被用于语音识别。具体实现时，归一化能量可作为 CEP 系数的维一分量，能量信息的加入对识别系统的性能有一定的改善[Rabiner 84, Furui 88]。

§2.2.3 倒频谱线性回归系数及距离度量

近年来的研究表明，在非特定人的识别中仅用倒频谱这种静态信息是远远不够的，为了减少不同说话人之间的差虽对识别系统的影响，不少学者提出用动态信息来加强特征表示。当不同的说话者发同一个音时，也许共振峰的位置发生了变化，但谱的包络线及能量的变化曲线大致不会改变，因此这些动态信息能从另一个角度表征语音。将语音的静态信息与动态信息一起作为语音识别的特征，取得了较好的效果[Furui 86, Soong 86, Lee 88b]。

常用的动态特征主要有倒频谱线性回归系数[Furui 86],或称为 Delta Cepstrum[Soong 86, Rabiner 88]：

$$R_m(t) = \left(\sum_{n=-k}^{k} n \cdot C_m(t+n) \right) / \left(\sum_{n=-k}^{k} n^2 \right)$$

和倒频谱差分系数[Paul 96, Lee 88a]。

$$D_m(t) = C_m(t+k) - C_m(t-k)$$

这里 $C_m(t)$ 是发音的第 t 帧 CEP 系数的第 m 阶系数，一般 k=(20～50ms)×采样率。

除此之外，类似的能量线性回归系数和能量差分系数也是很重要的动态信息。

静态和动态特征综合在一起作为语音识别的特征，如何度量距离成为一个主要的问题。一般来说，是用加权的距离度量：

$$D_{cep} = \sum_{m=1}^{ND}\left(C_m(i) - C_m(j)\right)^2 + W_d \sum_{m=1}^{ND}\left(C_m(i) - C_m(j)\right)^2 + W_p\left(C_o(i) - C_o(j)\right)^2 + W_q\left(D_o(i) - D_o(j)\right)^2$$

其中 $C(i)$是 LPC-CEP 系数，$D(i)$是倒频谱的线性回归或差分系数，C_0 是规一化能量，D_0 是能量的线性回归或差分系数。W_d、W_p、W_q 分别是加权因子，它们的值完全靠实验来确定。式中用的是欧氏距离，也可以用其它距离量度方法。这种综合特征及距离度量取得了较好的效果[Shikano 86, Tohkura 86]。

§2.3 聚类技术研究

聚类(Cluster)就是把一个 ND 维的欧氏矢时空间划分为 M 个区域，这 M 个区域分别由其中心矢量表征。这个过程需要一个由大量的矢量构成的样本集，经过统计实验后确定出 M 个中心矢量，这一过程叫做“训练”或“建立码本”，也就是我们所说的聚类过程。这 M 个中心矢量通常称为一个大小为 M 的码本(CodeBook)，每个中心矢量都称为一个码字(CodeWord).

聚类是矢量量化(Vector Quantization)技术首先要解决的问题。而聚类有一个重要的问题，就是如何确定一个准则，使得在这种准则下聚类过程达到最优，也就是用这 M 个中心矢量可以“最好”地表征这个样本集(样本空间)。

设有样本集

$$X = \{x_i\}, \quad 1 \le i \le N$$

(在这里，矢量 x_i 为 16 维 CEP 向量)，我们要把它聚成 M 类，

$$X = C_1 \cup C_2 \cup \cdots \cup C_M \ (C_i \cap C_j = \phi, \text{当} i \ne j)$$

这个分类记作 C，而其准则度量(距离)记作 $D(C)$。聚类的任务是对 X 作一个最准分类 $C^* = C_1^* \cup C_2^* \cdots \cup C_M^*$，使得 $D(C^*) = \min D(C)$。

一个理想的准则度量必须具备主观上有意义而且易于运算的特性，它有很多种，常用的有以下几种(其中 T、W 和 B 分别表示总的、类内和类间协方差矩阵)：

$$D_1(C) = tr(W) \quad \to Min$$

$$D_2(C) = tr(W^{-1}B) \to Max$$

$$D_3(C) = tr(T^{-1}B) \to Max$$

$$D_4(C) = \det W \quad \to Min$$

其中以 D_1 最为常用，称为最小平方距离准则。它可改写为：

$$D_1(C)=\sum_{j=1}^{M}\sum_{X_i\in C_j}\left\|X_i-\overline{X}(j)\right\|^2\rightarrow Min$$

或

$$D_1(C)=\sum_{j=1}^{M}\sum_{X_i\in C_j}\sum_{X_k\in C_j}\left\|X_i-X_j\right\|^2\rightarrow Min,\quad k>1$$

当把样本的欧氏距离项$\|\cdot\|^2$换为抽象距离，还可有不同的距离定义(加权距离、Mahalabis 距离，伪距离等)。

聚类方法很多，有 K-均值、ISODATA 等，下面将介绍三种不同聚类方法。

§2.3.1 K-均值法[Furui 89]

这是一种递归的聚类算法，它把训练矢量集$\{X_i\}$聚成 K 类$\{C_i\}$，有下面四个步骤：

步骤 1：初始化

设递归深度 m=0。用一种适当的方法选一个初始码本矢量集合$\{Y_i^{(m)}\}(1\le i\le K)$。

步骤 2：分类

按最近邻(Nearest Neighbour)准则把训练矢量集中的矢量 X 分到各某中：

$X\in C_i^{(m)}$ iff $d[X,Y_i^{(m)}]\le d[X,Y_j^{(m)}]$, 对所有 $j\neq i$

步骤 3：产生新码本

令 $m\leftarrow m+1$。计算每一类的质心

$Y_i^{(m)}=centroid(C_i^{(m-1)}),\quad 1\le i\le K$

以此作为新的码本矢量，并计算所有训练矢量的总失真度 $D^{(m)}$。

步骤 4：结束判断

如果 $D^{(m)}$ 比 $D^{(m-1)}$ 下降百分比达到某一阈值则停止，否则转步骤 2。

这里 $d[\cdot]$是样本之间的某种距离度量。

这种方法可以通过修改码本降低总失真度。但它有时可能收敛到一个比全局最优点差得多的局部最优点。它很大程度上取决于初始码本的位置，因此可以用不同的初始值不断聚类，并从中得到一个总失真度最低的(最好的)码本。

§2.3.2 LBG 算法

这是由 Standford 大学的 Linde, Buzo 和 Gray 等人提出的一种聚类算法[Linde 80]，因此人们泛称这种码本生成算法为 LBG 算法。

这种算法一般假定码本大小固定，而且为 2 的幂。码本开始很小，然后不断扩大，直到达到要求。它常把一个已存在的分类分裂成两个小类，并给每个小类以新的码字初值。

下面是其步骤[Furui 89]：

步骤 1：将整个训练集作为一个初始类。初始码本只有一个码字，即整个训练集的质心。如图 2.3a。

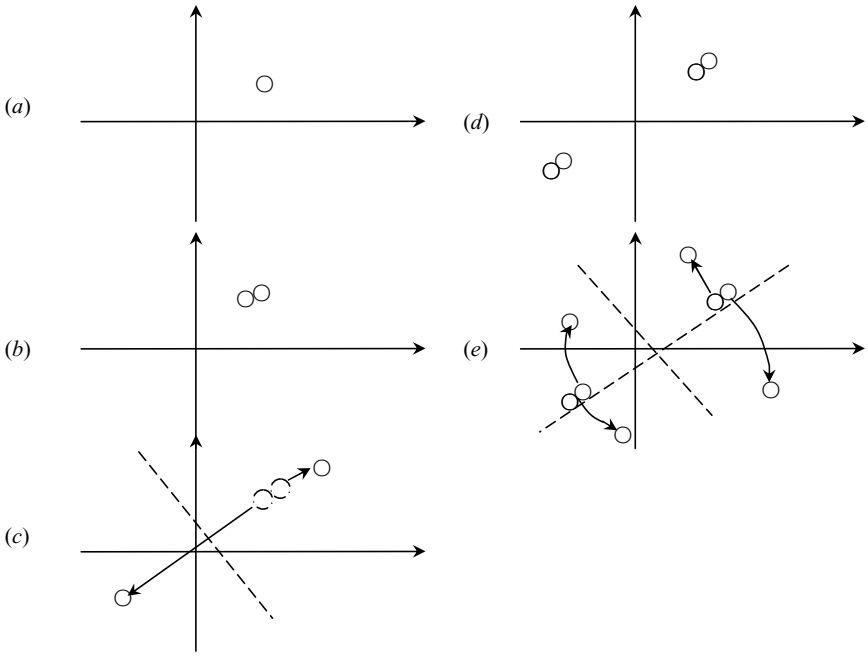

图 2.3 两维情形下 LBG 算法图示：(*a*)整个训练集的中心。(*b*)把该唯一码字分裂成两个初始码字估计。(*c*)把训练集按这两个码字聚类形成两个更好的码字。(*d*)把这两个码字分裂形成四个初始码字估计。(*e*)把训练集按这四个码字聚类形成四个更好的码字。

步骤 2：将该类分裂为两个子类，结果码本大小增大一倍，如图 2.3b、c。

步骤 3：重复这种“聚类—分裂”过程，直到码本大小达到要求，如图 2.3d、e。

我们把这一算法进一步完善，得到下面更详细的步骤：

步骤 1：初始化

设置迭代深度 m=0。把整个训练集作为一类，其质心作为初始码本的唯一码字，

$$B^{(m)} = \left\{C_i^{(m)}\right\}, \quad 1 \le i \le 2^m$$

步骤 2：分裂

将码本中每个码字 $C_i^{(m)}$ 作一小的扰动 Δi (Δi 为一小的扰动矢量)，得到一个新的码本：

$$B^{(m)} = \left\{C_i^{(m)} + \Delta i, C_i^{(m)} - \Delta i\right\}, \quad 1 \le i \le 2^m$$

步骤 3：聚类

将所有训练集按 $B^{(m)}$ 中码字重新聚类，并计算每一类的质心 C_i 作为新的码字

$$C_i^{(m)} = centroid(C_i^{(m)}), \quad 1 \le i \le 2^m$$

从而得到新码本：

$$B^{(m)} = \left\{C_i^{(m)}\right\},$$

步骤 4：失真度判断

计算总的失真度，当其下降百分比达到给定的阈值时转步骤 5；否则转步骤 3。

步骤 5：结束判断

令 $m \leftarrow m+1$。重复步骤 2、3、4 直到码本大小达到要求。

这种方法能得到较好的码本，但聚类时间却比较长。

§2.3.3 模拟退火 K-均值算法

徐雷在 89 年 3 月的《模式识别与人工智能》上介绍了一种新的聚类分析算法[Xu 89]，这一算法已由本教研组李建民、赵彤青及我等研制的语音识别系统中采用，取得了很好效果，使得聚类的时间大大降低[Zhao 90]。

在物理中，退火技术是指：先把固体加热至足够高，使固体中所有粒子处于自由液态，然后将温度缓慢下降，这样只要温度上升得足够高，冷却过程足够慢，则所有粒子最终会处于最低能态。

如图 2.4 所示，若粒子开始处于 C 状态，若让能量逐渐减小，则粒子最终到达的是 A 点(局部最低点)而不是 B(全局最低点)点，这是我们所不希望的。解决的办法是对系统经常地摇动一下，就很可能把粒子从 C 点摇到 B 点，而把它摇到 A 点的可能性很小。若开始以较大的速度摇，再慢慢减速，最终粒子会落在 B 点，模拟退火(Simulated Annealing)就类似于这个过程[Kirkpatric 83]。

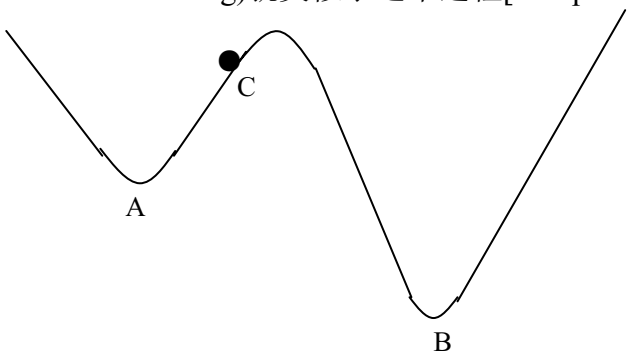

图 2.4 粒子在多能量极小点系统中的运行

在给定温度 T 下，处于热平衡状态的物体内能 E 服从玻尔兹曼(Boltzmann)分布：

$$P(T) = C(T)\exp(-E/(KT))$$

其中 K 为 Boltzmann 常数，$C(T)$为规一化因子。

Metropolis 抽样模拟了温度 T 下的热平衡：随机地选一个初始状态$\{X_i\}$，然后随机地给系统一个小扰动$\{\Delta X_i\}$，若内能增量$\Delta E < 0$，此扰动被接受；否则，此扰动以概率 $\exp(-\Delta E/(KT))$ 被接受。若扰动被接受，则系统状态$\{X_i\}$被$\{X_i + \Delta X_i\}$代替；否则产生一个新的扰动，…，如此下去直至$\{X_i\}$状态序列满足上式的分布。进一步地，让温度 T 从足够高的值逐渐下降，对每一温度 T，用

Metropolis 抽样使状态达到热平衡，一直到 T=0，此时物体达到 Ground 态，即 E 的最小值。

模拟退火 K-均值(ALK-means)利用退火原理及 K-均值法，其步骤如下：

步骤 1：随机分类 $C=\{C_1 \cup C_2 \cup \cdots \cup C_M\}$，并计算其总失真度 $d=d(C)$。设置初始温度 $T^{(o)}$，温度下降因子 $\lambda(\lambda<1)$。K=0，I=0。

步骤 2：$K \leftarrow K+1$，若 $K>N$ (样本总数)，则 $K \leftarrow K-N$。

步骤 3：若 $T<T\min$ (给定阈值)，结束。

步骤 4：对样本 $X_k \in C_i$，产生[1，M]上的随机整数 $j(j \neq i)$,并计算把 X_k 从 C_i 移到 C_j 后的总失真度的变化量

$$\Delta d=\left\|X_k-\overline{X}_j\right\|^2 \cdot \frac{m_j}{m_j+1}-\left\|X_k-\overline{X}_i\right\|^2 \cdot \frac{m_i}{m_i-1}$$

其中，m_i、m_j 分别为 C_i、C_j 中的元素个数，$\overline{X}_i$ 、$\overline{X}_j$ 分别为 C_i、C_j 中元素的均值。

步骤 5：若 $\Delta d \leq 0$：把 X_k 从 C_i 移到 C_j，I=0，$d \leftarrow d+\Delta d$，修改 m_i、m_j 及 $\overline{X}_i$ 、$\overline{X}_j$。转步聚 2。

步骤 6：产生[0，1]上的随机浮点数 ξ，若

$$\exp(-\Delta d/T) \leq \xi$$

则 $I \leftarrow I+1$。若 $I>IGM$ (给定阈值)，则 $T \leftarrow \lambda T, I \leftarrow 0$。转步骤 2。

以往的聚类算法，如 K-均值、ISODATA 算法等，都是组合优化的近似算法，只能求得局部最优解，而模拟退火 K-均值算法可以以很高的概率收敛于全局最优解，且与初始分类无关。

§2.4 矢量量化技术

矢量量化(Vector Quantization)技术是语音处理中最重要、运用最普遍的方法之一。假定 X 是一个 K 维向量，其各维分量都是实值随机变量。在 VQ 中，向量 X 要映射成另一个 K 维向量 Y，这称作把 X “量化”成 Y，写作：

$$Y=VQ(X)$$

Y 在一个有限集中取值，这个有限集就是一个码本，我们记作 $CB=\{CW_i\}$，$1 \leq i \leq NC$，NC 为码本大小。显然，VQ 的过程就是样本空间 X 到有限空间 CB 的映射：

$$x \in X \subset E^K \rightarrow Y=VQ(x) \in CB \subset E^K$$

当把 X 量化为 Y 后，它们之间存在一个量化失真或称距离度量 $d(X,Y)$。一个量化器 VQ(·)称为最优的是说它是所有量化器中平均量化失真

$$D=\sum_{x \in X} d(x,VQ(x))/\left|X\right|$$

最小的，其中 $\left|X\right|$ 表示集合 X 中元素的个数。

当我们选用下面形式的量化器(NN 原则)时，

$$Y = \mathrm{VQ}(X) = CW_i \quad \text{iff} \quad d(X, CW_i) < d(X, CW_j), \quad \text{对所有} j \neq i$$

对某一样本集 X 来说，一个(大小一定的)码本和一个量化器是一一对应的。

码本的选择由聚类算法实现，而 VQ 过程由最近邻(NN)原则很容易实现。一般来讲，特定领域(如语音领域)的样本(训练)集 X 是 E_k 的一个真子集，显然 X 越大、覆盖面越宽(相对于该领域)，对聚类越有帮助。

在实际的实现中，某一向量 X 对某一码本 CB 量化成 CW_i 后，为运算方便，只用该码字在 CB 中的编号 i 来表示量化结果。这样，VQ 可以表示为：

$$Y = VQ(X) = i \quad \text{iff} \quad d(X, CW_i) < d(X, CW_j), \quad \text{对所有} j \neq i$$

§2.5 基音检测的方法与研究

汉字的声调信息对基于音节的汉语语音识别和理解起着重要作用，是汉语语音识别的一个重要特征。而基音周期的变化是进行声调评价的一个有效方法。

基音周期的检测方法很多。频域的有基于短时傅里叶分析和同态语音处理的，时域有利用并联处理方法和自关函数的[Rabiner 78,Furui 89]。本文将采用用自关函数进行基音周期检测的方法。

从某种意义上说，自关表示法的一个主要限制是它保留的语音信号的信息太多，由于声道响应的阻尼振荡可能使自关函数有许多峰，如果窗选得比基音周期短，或共振峰频率快速改变，会使得自关函数不在基音周期附近取得最大值，从而导致简单地选取自关函数中最大峰的方法失败。

为解决这一问题，首先要确定合适的窗宽。考虑到人类语音基频范围一般为 80～500Hz，基音周期的变化范围为 2～12.5ms，因此选取帧长(窗宽)为 256 点(9.6KHz 采样)就可保证在任何情况下，一帧至少覆盖二个基音周期。

其次需要对语音信号加以处理。使周期性变化更加明显，同时抑止信号中其它带来扰乱的特征，这种“谱平整”[Rabiner 78]技术的主要任务是去掉声道转移函数的影响，从而使每个谐波有相同的幅度。其中较为有效的是“中心削波”技术。

§2.5.1 中心削波

中心削波是一个非线性变换：

$$Y(x) = C[x] = \begin{cases} x - C_L, & x \geq +C_L \\ x + C_L, & x \leq -C_L \\ 0, & \text{其它} \end{cases}$$

图 2.5 给出中心削波函数的波形及其作用于语音信号波形的一个例子：

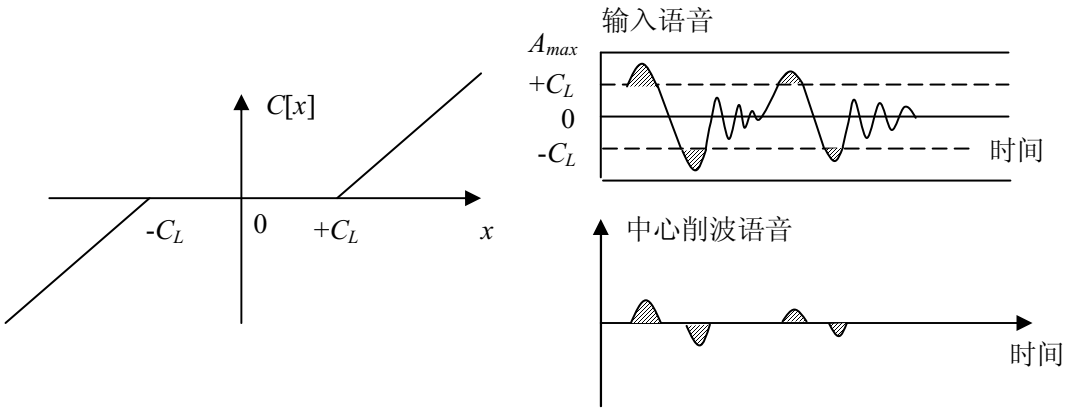

图 2.5　中心削波函数及其作用

Sondhi 用最大幅度值的 30%($A_{\max}\cdot 30\%$)作为 C_L 的值，Rabiner 建议用第一个 100 个抽样中最大值的 68%($A_{\max}\cdot 68\%$)作为 C_L，由 Rabiner 给出的实验结果我们可以看到，C_L 值越大越能削弱干扰引起的外来峰，使周期性越明显，如图 2.6 所示[Rabiner 78]。在我们的系统中我们用一帧内的最大幅度值的 68%作为 C_L。

§2.5.2　基音检测

经过处理后的语音信号就可以计算自相关函数。

$$R_n(k)=\sum_{i=0}^{N-1-k}S_n(i)S_n(i+k)$$

事实上相隔 1～20 点(20ms)以内的自相关函数不必计算。计算后的自相关函数按最大值 $R_n(0)$ 进行归一。从中找出所有非零区域的最大值点(其值超过 $R_n(0)$ 的 20%) $m_o,\cdots,m_t$，它们作为基音周期的候选。如何确定哪个 m_i 体作为基音周期？

为此首先定义近模函数 near_mod(*x*,*y*)为 *x* 离 *y* 的整数倍的最近距离。如：

near_mod(21,10)=1　　(21%10)

near_mod(28,10)=2　　(而不是 8)

其次定义评价函数

$$dis(i)=\sum_{j=i+1}^{t}near\_\mathrm{mod}(m_j,m_i)/(t-i)$$

则具有最小 *dis* 值的 m_i 作为理想基音周期。

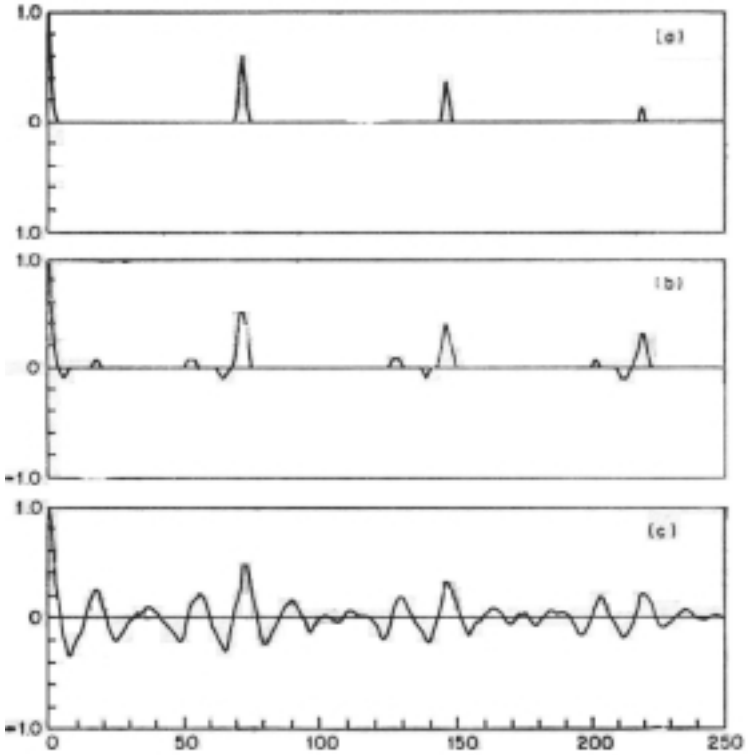

图 2.6 某段浊音的中心削波语音的自关函数

C_L 为最大值的 (a) 80% (b) 64% (c) 48%

例：m1=10, m2=22, m3=25, m4=29

则 dis(1) = (2+5+1)/3 = 8/3

dis(2) = (3+7)/2 = 5

dis(3) = 4/1 = 4

因此 pitch=m1=10。这种方法可以有效地滤掉那么像 m3=25 的不速之客。

当然越小的数越容易被判为基音周期的理想值，这是显而易见的。但对特定问题情况却不同。人的基频是 70～500Hz，基音周期即为 14.3～2ms，对 9.6KHz 采样相当于 140～20 个点。有了这个知识再加上自相关函数幅度(百分比)的约束，可有效防止意外情况。

在帧移 128 点(半窗宽)情况下，下帧的基音周期不会发生突变，因此只需在上一帧基音周期估值附近(如±20 个点范围)寻找，这样就可得到一个基音周期估值序列：

$$P_1, P_2, \cdots, P_n$$

以它作为声调评价依据。这一序列一般不会有倍周期，但可能会有分周期，为此统计出概率分布最密点的值 $P_{\max}$，将这一序列中 $[P_{\max}/2, P_{\max}*2]$ 范围外的以及突变的点去掉，用插值补充。

§2.5.3 声调评价

对于一个数字(0～9)识别系统，可以把它分为三类：一声类(1，3，7，8)；

四声类(2，4，6)和二、三声类(0，5，9)。实验表明，这种分法具有类内可分性。

我们从基音周期估值序列$\{P_i\}$中抽出两点(60%和 90%处的三点或五点平滑值)P_b和P_e，利用两点比较法即可以判出三类声调来。

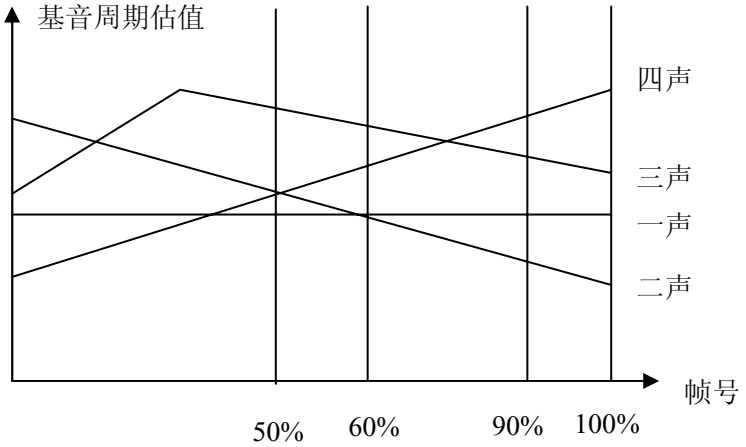

图 2.7　汉语四个声调的基音周期走向图

当然为了节省时间，只需在语音的后一半中进行基音周期检测，由图 2.7 可以看出，这样并不影响三类声调的判别。这种两点比较法有效地避免了基频的“弯头”和“降尾”的影响。

由有这种方法判断出三类声调后，这种结果以概率的形式参加总体的识别，而不是以粗分类的方式参加进去，这样防止了由粗分类的错误剪枝带来的识别率下降。实验表明，虽然数字发音会出现变调情况(如 595 发成为 w ú　j í u w ǔ)，虽然声调的识别也会发生错误，但总体识别率比每一种(单音节、声调)识别率都要高。

第三章　隐马尔可夫模型的语音识别中的应用

隐式马尔可夫模型(Hidden Markov Model)最开始出现在 Baum 等人的文章[Baum 72]中，紧接其后，它分别被 CMU 的 Baker 等人[Baker 75]和 IBM 的 Bakis、Jelink 等人[Bakis 76, Jelink 76]引入语音识别领域。在八十年代初美国 Bell Lab 的 Rabiner 等人提出了这一方法用于非特定人的语音识别[Rabiner 83]。短短的十几年时间，HMM 成为语音识别中一种很有效的技术，它不仅能用来作为(以音素、音节或词为单位的)语音产生的声学模型，而且能作为词法、语法、语义等高层次的语言模型，因此 HMM 在语音识别领域得到了广泛的应用，也取得了很多成就。

由于 HMM 是一种随机概率模型，因而它能较好地瓜人类发音过程的变化规律，模型也稳定得多。虽然它的训练过程比较复杂，但模型所占的存储空间较传统的方法少，作分类比较的返算也比较快，故而识别系统简单得多。HMM 方法的出现极大地推动了非特定人语音识别的发展。

§3.1　马尔可夫过程及马尔可夫链简介[Lu 86]

马尔可夫过程是一个随机过程$\{\xi(t), t\in T\}$，它具备这样的性质，即可知t_m时刻过程处在状态$\xi(t_m)=i_m$的条件下，在时刻t_m以后过程将要到达的状态的情况与该时刻以前过程所处的状态无关。这个性质也称为过程的无后效性或过程的马尔可夫性。

马尔可夫过程$\{\xi(t), t\in T\}$可能取的值的全体构成过程的状态空间。状态空间可以是连续的，也可以是离散的；马尔可夫过程的参数 t 可以是连续的，也可以是离散的。

对一个状态空间(I)离散、参数为非负整数的随机过程$\{\xi(t), t=0,1,2,\cdots\}$，若它满足条件。

$$\mathrm{Prob}\{\xi(t+1)=j|\xi(0)=i_o,\cdots,\xi(t)=i_t\}=\mathrm{Prob}\{\xi(t+1)=j|\xi(t)=i_t\}$$

这样的随机过程称为马尔可夫链。马尔可夫链在 t 时刻的一步条件转移概率

$$a_{ij}(t)=\mathrm{Prob}\{\xi(t+1)=j|\xi(t)=i\}$$

也称作 t 时刻的状态转移$(i\to j)$概率。显然有：

$$a_{ij}(t)\geq 0,\qquad i,j\in I$$

$$\sum_{j\in I}a_{ij}(t)=1,\qquad i\in I$$

§3.2　HMM 用于语音识别的基本原理及其定义

一个隐马尔可夫过程是由两个相互联系的过程(一个状态空间有限的马尔可夫链和一个随机函数集)作用而产生的随机过程。随机函数集中每一个元素都与

某个状态相关联。在时间离散的条件下，假定整个过程处于马尔可夫链的某一状态，其观察结果是与当前状态有关的随机函数产生。马尔可夫链的状态转移取决于状态转移概率矩阵。观察者只能看到与每一状态相关联的随机函数的输出，而不能看到马尔可夫链的状态变化，这就是所谓“隐”的含义。

用 HMM 原理进行语音识别是基于这样的一种想法：从人类的发音机理出发，声门激励的形式和声道的形状作为发音动作结构(或称状态)决定了所发的声音。而从工程角度认为某种语言只对应有限个发音动作结构(或称状态)是合理的。每一种动作结构可产生相应的短时信号，若干个短时信号的连接就组成了一段语音。这段语音中短时信号的变化反应了发音动作结构(或称状态)的变化。这种状态的变化过程符合马尔可夫过程的变化规律。

离散 HMM 定义如下[Furui 89]，其中观察符号是量化的随机函数值的符号表示：

$Q=\{q_1,\cdots q_N\}$: 模型中的(隐)状态集

N=状态数目

$V=\{v_1,\cdots,v_M\}$：观察符号集合(VQ 码本)

M=观察符号数(VQ 码本大小)

$A=\{a_{ij}\}$，状态转移概率矩阵

$$a_{ij}=\mathrm{Prob}(q_j\ at\ t+1|q_i\ at\ t)$$

$B=\{b_j(k)\}$，在状态 j 下观察符号的输出概率矩阵

$$b_j(k)=\mathrm{Prob}(v_k\ at\ t|q_j\ at\ t)$$

$\pi=\{\pi_i\}$，初始状态矢量

$$\pi_i=\mathrm{Prob}(q_i\ at\ t=1)$$

确定一个 HMM 模型需要确定状态数 N，离散符号数 M(码本大小)，以及三个概率密度 A，B 和 π。因此 HMM 模型常用 $\lambda=(A,B,\pi)$ 来简记。

在给定了一个 HMM 模型后，观察序列 $O=\{0_1,\cdots,0_T\}$ 的产生如下，

步骤 1：设置 t=1。

步骤 2：根据初始状态分布 π 选择初始状态 i。

步骤 3：根据 i 状态下的符号概率分布 $b_i(k)$ 选择 0_t。

步骤 4：根据 j 状态下的状态转移概率分布 $\{a_{ij}:j=1,2,\cdots,N\}$ 选择 j。

步骤 5：$t\leftarrow t+1, i\leftarrow j$。如果 $t<T$，转步骤 3；否则过程结束。

§3.3 HMM 的三个关键问题

当使用 HMM 模型时，有三个关键的问题要解决：

问题 1：Training Problem

给定许多观察序列 $O=\{0_1,\cdots,0_T\}$ 如何调整模型的参数 $\lambda=(A,B,\pi)$，以使 $P(O|\lambda)$ 最大？

问题 2：Evaluation Problem

给定一个观察序列 $O=\{0_1,\cdots,0_T\}$ 和模型 $\lambda=(A,B,\pi)$，如何计算概率 $P(O|\lambda)$？

问题 3：Hidden State Sequence Uncovering Problem

给定一个观察序列 $O=\{0_1,\cdots,0_T\}$，在某种意义下，如何选择一个最优的状态序列 $I=\{i_1,\cdots,i_T\}$？

第一个问题的解决用于获得 HMM 模型的参数，以便建立模型。第二个问题的解决可以用于根据观察序列计算每个模型的得分从而实现对未知语音的识别。第三个问题的解决可以使 HMM 在连续语音识别中发挥作用。

第一个问题的解决用通常被称为“向前—向后”算法(Forward—Backward)的 Baum-Welch 算法[Baum 72]解决。对语音观察序列 $O=\{0_1,\cdots,0_T\}$，引入向前概率函数：

$$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)\,a_{ij}\right]b_j(0_{t+1}),\quad 1\le t\le T-1$$

及

$$\alpha_1(i)=\pi_i b_i(0_1),\quad 1\le i\le N$$

同时引入向后概率函数：

$$\beta_t(i)=\left[\sum_{j=1}^{N}a_{ij}b_j(0_{t+1})\right]\beta_{t+1}(j),\quad 1\le t\le T-1$$

及

$$\beta_T(j)=1,\quad 1\le j\le N$$

这样模型的参数计算的迭代公式为：

$$a'_{ij}=\frac{\sum_{1\le t\le T-1}\alpha_t(i)a_{ij}b_j(0_{t+1})\beta_{t+1}(j)}{\sum_{1\le t\le T-1}\alpha_t(i)\beta_t(i)}$$

$$b'_j(k)=\frac{\sum_{1\le t\le T,\,0_t=V_k}\alpha_t(j)\beta_t(j)}{\sum_{1\le t\le T}\alpha_t(j)\beta_t(j)}$$

$$\pi'_i=\frac{\alpha_1(i)\beta_1(i)}{\sum_{1\le i\le N}\alpha_T(i)}$$

这个问题解决以后，第二个问题很容易得到解决。模型产生观察序列的概率为：

$$P=\sum_{1\le i\le N}\alpha_T(i)$$

第三个问题由 Viterbi 算法[Viterbi 67]解决。

令

$$\phi_1(i)=\pi_i b_i(0_1),\quad 1\le i\le N$$

进行如下递归 $(2\le t\le T,1\le j\le N)$

$$\phi_t(j) = \max_{1\le i\le N}[\phi_{t-1}(i)a_{ij}]b_j(0_t)$$

及

$$\psi_t(j) = i^{\max}\ (\text{使}\phi_{t-1}(i)\text{最大的}\ i\ \text{值})$$

最后得到

$$q_T = i^{\max}\ (\text{使}\phi_T(i)\text{最大的}\ i\ \text{值})$$

$$q_{t-1} = \psi_t(q_t),\quad 2 \le t \le T$$

而产生这个状态序列的概率为：

$$P = \phi_T(i^{\max})$$

§3.4 实际应用中几个需要注意的问题

§3.4.1 模型的参数初始化

Baum-Welch 算法是收敛的，但它却存在一个问题：它只能收敛到局部最优，而不可能收敛到全局最优，因此模型参数的初始化显得很重要，遗憾的是至今还没有人能有效地解决这一问题。比较常用的简单的方法是等概率初值的方法[Rabiner 83]。Juang 等人给出一种较复杂的 Segmental K-means 方法，它给出了较为精确的初始值[Juang 85a, Rabiner 85]。

Lee 却认为没有必要对离散参数的 HMM 给予太复杂的初值[Lee 88a]，只需一个统一的分布即可。但对连续参数的 HMM，或参数过多时，模型的性能对初值很敏感，此时需给出较为恰当的初值[Paul 88]。

§3.4.2 输出概率矩阵的平滑问题

训练集的有限性，使得训练完成后 B 矩阵中有一些零元素(零概率)，这些不合理的零概率给识别带来一定的影响。解决这个问题的最简单的方法是 Floor method [Lee 88a]，它将 B 矩阵的零元素赋给一个最小值 ε [Levinson 83]，然后修改 B 矩阵的其他元素使它满足约束条件。

这种方法简单、方便，但却不能区分“不容易”出现的码字之间的差别。改进的方法有 Distance Method，它的基本思路是根据码字之间的距离对 B 矩阵进行平滑：如果一个零概率码字与一个很高概率的码字很近，那么它的概率将得到显著的提高[Cravero 84, Schwartz 84]。

另一种方法称为 Co-occurrence Method [Sugawara 85, Lee 88a]。它的基本思路是根据码字之间容易相互替代的信息进行平滑：如果一个零概率码字在一定上下文环境中很容易被另外的非零概率码字所代替，那么它的概率将根据这个非零概率值加以修正。这种信息通过同一发音的 DTW 匹配可以得到，也可以直接从输出概率矩阵中得到。

§3.4.3 运算的下溢

在使用“向前—向后”算法进行训练的过程中，随着 t 的增大，$\alpha_t(i)$、$\beta_t(i)$ 的值将很快接近零，出现下溢，这是因为计算机的字长有限。为了解决这个问题，对 $\alpha_t(i)$、$\beta_t(i)$ 乘上与 i 无关的因子，在修正 a、b 参数时分子分母同时把因子约去[Levison 83]。这个加权因子一般取：

$$C_t = \left[\sum_{i=1}^{N} \alpha_t(i)\right]^{-1}$$

另一种方法是采用对数压缩方法[Brown 87]，将概率对数化并用定点数表示，这样做不但解决了这一问题，而且还减小了存储量、缩短了运算时间。

§3.4.4 HMM 状态数目和模型结构

Rabiner 等人的实验表明，状态数目超过 5 对识别率没有改善[Rabiner 83]。[Jiang 89]的实验也认为具有 5～6 个状态的 HMM 对孤立词的识别已足够了。

关于 HMM 的模型结构，对孤立词的识别来说，一般采用“从左到右”模型。普通的 HMM 被认为是全状态转移的，但我们希望在应用中对状态转移加以适当的限制。通常把状态转移矩阵 A 限制为上三角的，这样状态转移只能发生在 S_i 到 S_j 之间($i \le j$)，这样的 HMM 模型称为是“从左到右”的模型。这种模型的拓扑结构包含了时间信息，因为前面状态的输出观察值必定在后面状态的输出观察值之前，于是使得模型能适应语音的时序性。

最典型的 HMM 模型结构有下面两种：

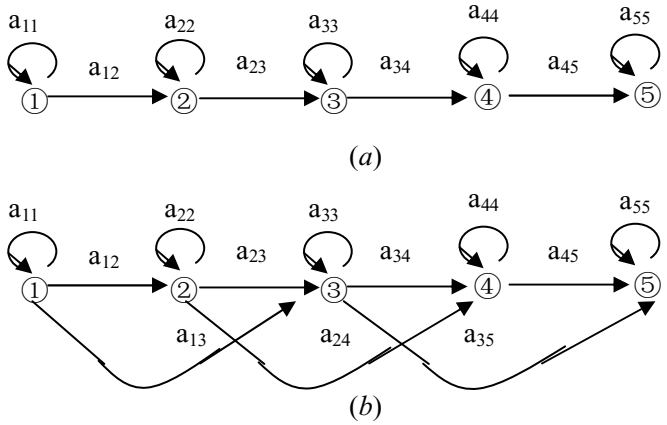

图 3.1 两种典型的“从左到右”HMM 模型结构

对于“从左到右”模型，只要稍稍修改训练算法即可，这种模型的初始状态始终在第一个状态，并且认为多套训练样本是相互独立的。

[Jiang 89]的实验表明这两种结构的 HMM 的性能没有显著的差别。Rabiner 等人还比较了并行结构的 HMM 与串行结构 HMM 的差别，在识别性能上发现它们没有明显的不同。因此我们认为，对孤立词的识别来说，图 3.1(a)结构的 HMM

是非常合适的。

§3.4.5 模型的训练算法

Baum-Welch 训练法是基于最大似然估计(MLE)的，MLE 方法有一个根本的假设即语音观察序列确定由模型产生，即假定模型一定是正确的，但如果这个假设不成立，那么 MLE 方法也许达不到所期望的结果。Baul、Brown 等人提出的最大交互信息估计法(MMIE)就是针对这个问题的[Baul 86, Brown 87]，MMIE 估计模型参数的过程如下：首先用 MLE 方法训练得到模型，然后用这个模型对训练集进行识别，根据每一个发音对某个模型的概率修正这个模型的参数，修正过程迭代多次人至模型稳定即可。MMIE 从信息论的角度入手，充分考虑字表中各词的相互影响，虽然从理论上证明在模型正确的情况下它不可能比 MLE 更好[Nadas 83]，但在实际应用中，HMM 与语音的产生过程有一定的差距，因此 MMIE，也确实比 MLE 取得了更好的效果[Baul 86, Brown 87, Nadas 88]。这咱训练方法已被一些系统采用[Baul 88b, Merialdo 88]，并取得了很好的效果。

但用 MMIE 训练模型时计算量却非常大，尤其对于大字表语音识别系统。Baul 等提出一种新的训练算法称为 Corrective Training 算法[Baul 88a]虽然不能证明它收敛，但实验表明这种方法是有效的。与 MMIE 不同的是，Corrective Training 仅根据某一个发音的一些相近音对模型的概率来修正这个发音的模型参数。虽然这种方法比 MMIE 减少了许多计算量，但无论如何，这两种方法都需要在用 MLE 训练所得参数的基础上根据字表及训练集的识虽信息进行参数修正，比 MLE 方法多了几个“识别—修正”的迭代过程，计算量有了显著的增长。

§3.4.6 模型的识别算法

在基于 HMM 的孤立词语音识别系统中，Baum-Welch 算法[Baum 72, Levison 83]和 Viterbi 算法[Viterbi 67]都可以用来计算观察矢量对模型的概率，它们的性能大致相同[Babiner 83, Rabiner 86b]。前者把可能产生观察矢量的所有状态序列对应的概率都包括了，而后者则是求产生观察矢量的最大似然状态序列对应的概率，因而是一种动态匹配的过程，它在观察序列与模型状态之间求得一种最佳匹配。这两种方法的计算量大致相同，但由于 Viterbi 算法可将概率对数化并将浮点乘法化为定点加法，大大加快了识别速度，所以基于 HMM 的系统多采用 Viterbi 算法进行识别。

§3.5 连续参数的 HMM 与离散参数的 HMM

对于基于离散输出概率矩阵的 HMM 模型，VQ 量化误差及训练集的有限性都在某种程度上影响着系统的识别率。为解决这两个问题，连续参数的 HMM 被引入语音识别的领域。与离散参数条件下的有限观察矢量空间不同，连续参数的 HMM 假定观察矢量服从某一种分布，最常用的输出概率密度是多维的高斯概率

密度函数[Paul 86]，这种概率密度多用均值矢量和方差矩阵来表示。出于计算量的考虑，方差矩阵有时简化为对角阵。其他的分布还有高斯混合密度[Rabiner 86a]，自回归混合密度[Juang 85b]，拉普拉斯混合密度[Ney 88]等等。

采用连续参数 HMM 不但可以避免 VQ 引入量化误差，同时在一定的程度上减少参数，比离散参数 HMM 要求训练样本少。但目前还没有一种分布与语音特征非常吻合，而且无论训练还是识别，计算量都比离散参数情况下要多得多；而且连续参数的 HMM 对初值非常敏感。

对连续和离散参数 HMM 进行识别的性能评价，有两种截然不同的结果。Brown 对于产生这种结果的原因进行了分析[Brown 87]。对连续参数的 HMM 来说，必须保证对语音观察矢量分布的假定的正确的，当这个假设不成立时，模型训练通常所用的最大似然估计法 MLE 得到的参数可能令人失望；而对离散参数 HMM，只要训练数据充足，就能得到较好的统计值。他认为如果要用 MLE 估计连续参数的 HMM，则需要假定一个较精确的特征矢量分布。

基于将连续和离散参数的 HMM 结合起来的目的，黄学东等[Huang 89]提出了一种半连续马尔夫模型(Semi-continuous HMM)，这个模型的输出概率矩阵是离散的，但把 VQ 的码本看成模型的一部分，码本本身有一个分布，通过训练使模型和码本结合在一起达到最优化，量化时加上码字的分布信息，减少了量化误差对识别性能的影响，取得了比离散和连续参数 HMM 更好的效果。

第四章　基于非线性分块原理的分段概率模型

无论从理论还是从实践来看，HMM 用于语音识别都是一种很好的模型，尤其对于非特定人语音识别。但对于模型的结构窖是否符合人类语音产生的机理以它的结构和参数是如何影响识别性能的这两个问题，人们依然在探讨之中。严格地讲，只有时变的 HMM 才能更加准确地反映人类的发音的过程，而现有的 HMM 一个较明显的缺陷就是状态转移概率与时间无关，因此状态转移概率应随在状态的停留时间而变，而不应是一个确定的值。但这种模型的参数估计问题还没得到有效的解决，因此不少人试图在 HMM 思想的指导下用一些其它结构的模型来弥补这个不足。其中较有成效的是 Russell 等人提出的半马尔可夫模型(Semi-Markov Model)[Russell 85]，SMM 除了 HMM 的状态转移概率和输出概率以外，又加入了状态驻留概率(State Duration Probability)，取得较好的效果。Lee、Rabiner 等人先后在基于 HMM 的语音识别系统中加入状态驻留概率，以提高识别率。

另一些尝试是改造 HMM 以降低它的算法复杂性，同时使模型的状态转移概率不再与时间无关，这方面，茅为华、王作英、曹洪等人都作了不少的工作[Mao 88, Wang 88, Wang 89, Cao 89]，也取得了不错的效果。

蒋力提出了一种基于非线性分块(Non-Linear Partition)的概率模型 SPM(Segmental Probabi-listic Model)[Jiang 89]，实验表明，它不仅在算法复杂度上比 HMM 减少了几个数量级，而且取得了比 HMM 更好的识别率。作者在此基础，对算法作了一些改动，使得识别率进一步提高。

§4.1　NLP 原理

非线性分块(NLP)原理是这样一种算法，它根据语音特征信息的变化情况，将特征序列分为相对平稳的几块，从而可以起到压缩信息和时间规整的目的。对不同的发音来说，语音特征信息的变化情况在时间轴上的分布不同，但对同一发音来说都存在着较好的稳定性。这种方法取得了很好的效果[Jiang 89]。

设有语音观察序列 $O=\{C_1,\cdots,C_T\}$，其中 C_i 为 K 阶 CEP 系数矢量，$C_i=(C_i^{(1)},\cdots,C_i^{(K)})$，$T$ 为观察序列长度。要将观察序列分成 M 块，$O=P_1+P_2+\cdots+P_M$(其中“+”表示串接)，为此定义语音特征变化信息为：

$$d_j=d_{cep}(C_j,C_{j+1})$$

$$=\sum_{k=1}^{K}\left[W_k(C_j^{(k)})-C_{j+1}^{(k)}\right]^2,\quad 1\le j\le T-1$$

其中 $W=(W_1,\cdots,W_k)$ 是距离加权矢量。定义平均变化信息：

$$\Delta D=\frac{1}{M}\sum_{j=1}^{T-1}d_j$$

则当 $m_i(1\le i\le M,令 m_o=0)$ 满足下式时：

$$\sum_{j=1}^{m_{i-1}} d_j < i * \Delta D \le \sum_{j=1}^{m_i} d_j$$

以 CEP 系数序列 $C_{m_{i-1}+1} \sim C_{m_i}$ 作为第 i 块，记作 P_i，其长度 $m_i - m_{i-1}$，记作 L_i。

§4.2 SPM 原理

首先介绍[Mao 88]中提出的一个基于非线性分块原理的模型。类似于 HMM，它有输出概率矩阵

$$B = \{b_{ij}\}, \quad 1 \le i \le M, 1 \le j \le NC$$

其中 M 为模型状态数(即分块数)，模型的状态为

$$S = \{s_i\}, \quad 1 \le i \le M$$

NC 为码本长度，码本为

$$V = \{v_i\}, \quad 1 \le i \le NC$$

假设训练遍数为 L，则 B 矩阵的计算公式为：

$$b_{ij} = \sum_{k=1}^{L} \text{第}k\text{遍发音中}S_i\text{中出现码字}V_j\text{的个数}$$

在识别时，观察序列(码字序列) $O = \{C_1, \cdots, C_T\}$ 被分成 M 块 $\{P_i, 1 \le i \le M\}$，长度分别为 L_i。对待识字表 $\{w_1, \cdots, w_d\}$ 中每待识字 w(它对应的矩阵为) $B^{(w)} = (b_{ij}^{(w)})$ 统计分值：

$$F^{(w)} = \sum_{i=1}^{M} \left[\sum_{\substack{m_{i-1} < t \le m_i \\ C_t = V_j}} b_{ij}^{(w)} \right]$$

找出 $F^{(m)} = \max_{1 \le w \le d} \{F^{(w)}\}$，则认为识别结果为 W_m。

这个模型是一个计数模型，它虽能在一定程度上体现语音特征分布，却没有用到总体信息。$\{b_{ij}\}$只表示第 i 块中码字 V_j 出现的次数，而没有考虑其它码字的出现情况，即 $\sum_j b_{ij}$ 不规整，因此它不是严格意义上的 HMM 模型。

SPM 模型[Jiang 89]对 b_{ij} 的计算作了改进，认为

$$b_{ij} = \frac{\sum_{k=1}^{L} \text{第}k\text{遍发音中}S_i\text{中出现码字}V_j\text{的个数}}{\sum_{k=1}^{L} \text{第}k\text{遍发音中}S_i\text{中出现码字的总数}}$$

它保证了 $\sum_j b_{ij} = 1$。

识别时求出每个码字的计分值：

$$F^{(w)} = \prod_{i=1}^{M} \left[\prod_{\substack{m_{i-1} < t \le m_i \\ C_t = V_j}} b_{ij}^{(w)} \right]$$

找出 $F^{(m)}=\max_{1\le w\le d}\{F^{(w)}\}$，则认为识别结果为 w_m。

与 HMM 相比，这种模型仅有输出概率信息而无状态转移信息。它是一个概率统计模型，比[Mao 88]模型的效果要好得多，但它却没有考虑到分段的长度对识别的影响，可以想象，由于发音长度的不同，在 P_i 块中出现一个码字 V_j 与出现两个、三个以至多个 V_j 应该是相同的，但由于 b_{ij} 一般都小于 1，使得：

$$b_{ij}b_{ij}<b_{ij}$$

也就是说，发音越长，概率值越小，这显然有悖常理。为此我们对识别策略作了改进：

$$F^{(w)}=\prod_{i=1}^{M}\left[\prod_{\substack{m_{i-1}<t\le m_i\\ C_t=V_j}} b_{ij}^{(w)}\right]^{1/L_i}$$

其中 $L_i=m_i-m_{i-1}$。它在一定程度上对分块长度信息进行了综合。

在实际构造一个矩阵(或称模型)时，为了使 $b_{ij}\neq 0$，采用 Floor 方法设置最小值 ε，使得 $b_{ij}\ge\varepsilon$。

§4.3 SPM 中 ε 与状态数的选择

对离散参数的 HMM，为了解决有限训练集产生的 B 阵零元素，避免识别时产生不合理的零概率，因此将计算出的参数 $\{b_{ij}\}$ 中的零值赋于一个最小值 ε。

设 $B=\{b_{ij}\}$ 的第 i 行有 R_i 个零值，则作如下参数调整：

$$b_{ij}^{(a)}=\begin{cases}(1-R_i\varepsilon)b_{ij}, & b_{ij}\neq 0\\ \varepsilon, & b_{ij}=0\end{cases}$$

这样

$$\begin{aligned}\sum_j b_{ij}^{(a)}&=R_i\varepsilon+\sum_j(1-R_i\varepsilon)b_{ij}\\ &=R_i\varepsilon+(1-R_i\varepsilon)\sum_j b_{ij}\\ &=1\end{aligned}$$

一般认为 ε 值应选在 10^{-6}～10^{-4} 之间[Levinson 83]，而[Jiang 89]在对 DB30 语音数据库的 20 遍测试集的识别率(M 取 6)进行的实验得出如下结论：训练集识别率随 ε 值增大而下降，测试集识别率随 ε 的增加而上升(ε 在 10^{-8}～10^{-3})。这是由于有限训练集造成的。ε 的设置改变了原有训练参数，引起训练集识别率的下降；但对测试集， ε 越小则对测试集的适应能力越差。在我们的系统中选取 $\varepsilon=10^{-4}$。

SPM 的状态数(即非线性分块的块数)M 对识别率会有什么影呢？我们对二十六(‘0’~‘9’，‘05’~‘95’，‘11’，‘51’~‘91’)词表数据库进行实验(见第五章)认为状态数取 M=5 对数字识别来说是比较合适的。

§4.4 VQ 与分块的先后次序对识别率的影响

SPM 的关键在于平稳段的分割，为了了解分割点的位置对识别率的影，我们进行了以下实验。同样用 NLP 算法，但在帧与帧之间的距离度量上有所不同，一种是量化前的 CEP 系数的绝对值距离，另一种是量化后码字之间的距离(见第五章)。从实验可以看到，SPM 分块的稳定性对识别率有影，后者比前者识别率要低。

§4.5 聚类的类内离散度对识别率的影响

在用训练集中的 CEP 向量进行聚类时，我们发现，各类的类内离散度或类半径并不一样，因此在进行 VQ 时仅仅用最近邻(NN)原则是否可以？如图 4.1，两类的类中心分别为 A 和 B，类内样本到类中心的距离均值(我们称之为类半径)分别为 R_A 和 R_B，而且 $R_A<R_B$，样本空间内有一点(CEP 向量)P，它离 A 较 B 稍近，但从图中看，它应该量化成 B。基于这种考虑，在计算 P 到到各个类中心(码字)之间的距离时，用类半径进行规整。此时，距离公式为：

$$d(p,CW_i) = d_{cep}(p,CW_i) / R_{CW_i}$$

实验的结果(见第五章)并不象想象的那么好：男声并未改善，反而更差；女声和男女综合的模型的效果比原先好。这可能是由于女声发音的频谱特性没有男声的那么集中造成的；也有可能是因为聚类时作用使用的是 NN 原则的比重故，这有待进一步研究。但我们的这种初衷显然可以由学习矢量量化(LVQ)来解决[McDermott 89]。在实际的系统中，我们没有采用对距离进行规整的策略。

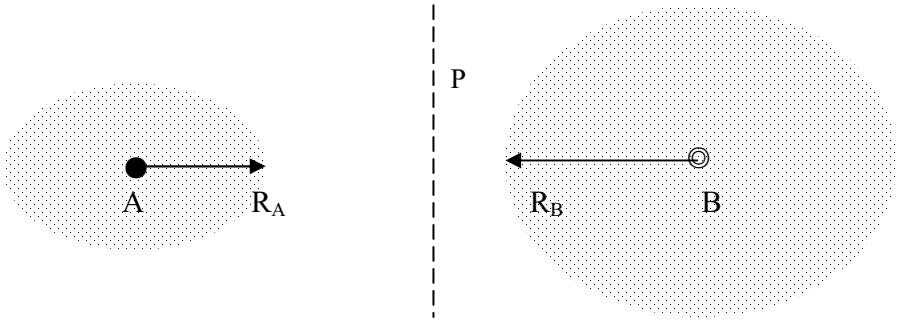

图 4.1 二维空间的 P 点的量化

§4.6 实际使用时的考虑

由于乘法、开方运算会降低运算效率，而且容易出现下溢，因此在实际识别时，模型中存的并不是概率值，而是概率的对数值(再乘以一个因子)，因而可使用定点数进行计算，令

$$Q(q_{ij}) = K \cdot \log_{10} B(b_{ij})$$

则识别计分公式可改为

$$F^{(w)}=\sum_{i=1}^{M}\frac{1}{L_i}\left[\sum_{\substack{m_{i-1}<t\le m_i\\ C_t=V_j}} q_{ij}^{(w)}\right]$$

使计算速度加快。

第五章　连续非特定人汉语数字识别系统的构成

在国内外都有人对数字识别方法进行了大量的研究。76 年 Rabiner 使用判定树的方法进行非特定人连接数字识别[Rabiner 76]，之后他又分别用连续概率分布 HMM 模型进行孤立数字识别[Rabiner 86a]，用 Delta-Cepstrum 的 HMM 模型乾地连续数字识别[Rabiner 88]，用含有状态驻留时间概率分布及能量概率分布的 HMM 模型进行了连续数字识别[Rabiner 89]，都取得了比较好的效果，但数字串的长度都不太长(＜8)，而且非特定人未知长度的数字串的识别率也不太高。1988 年高雨青使用倒谱回归系数为参数对二字数字串的非特定人连续数字识别方法进行了研究，未经训练的人测试识别率为 97.8%，经过训练的人测试识别率几乎为 100%[Gao 88]。1990 年苑宝生等使用并行快速动态时间弯折(PFDTW)方法实现了一个特定人口呼算题系统，平均(数字)识别率为 98.5%[Yuan 90]。本文将介绍一个数字识别系统，它使用分段概率模型进行非特定人随机长度(3～13)的连续数字识别。

本系统是非特定人连续数字识别系统，它的基本要求是对任何人连呼的至多 13 个数字的数字串进行识别。对 13 个连续数字识别来说，存在以下几个需要解决的问题：

(1) 数字串中各数字之间没有语法可依，也就是说没有知识可用于数字串的识别，因此要求音节识别率很高。

(2) 音节切分的正确性直接影响总体识别结果，这就对音节切分的正确性提出了很高的要求。

(3) 13 的数字的语音所占空间很大，而 TMS320C25 的片内数据存储器的容量远达不到要求，需要寻求解决这个问题的方法。

在第二章，我们已说明问题(2)的解决方法，问题(3)在一定的硬件配置下只需要设计一个好的算法就可以解决，问题(1)才是重要而且较为困难的。为此我们进行了大量的实验，旨在选择一种比较好的“识别”方法。

§5.1　实验及其结论

在进行实验时，我们定义了两个词表：词表 1 包括数字‘0’~‘9’；词表 2 除了词表 1 的内容之外，还包括了几组不易切分的二字串：‘05’~‘95’和‘11’，‘51’~‘91’，共 26 个字。

实验所用的数据来自十男十女所发的 20×5=100 遍语音。由这些样本中计算出的 CEP 矢量共 97615 个。我们所建立的模型相应为“男性”、“女性”和“混合”模型。

§5.1.1　CEP 权矢量的确定

我们发现，从语音数据中提取出来的 CEP 矢量各维的幅度并不均匀，为了使各

维对距离计算的贡献相差不至太悬殊，我们对 CEP 矢量进行了加权。权矢量的选择准则是：使用它之后，CEP 矢量的(统计)幅度均值应大致相等，(统计)方差也大致相等。为此，我们对 97615 个 CEP 矢量进行了统计，结果如表 5.1。

表 5.1　CEP 矢量的幅度均值、方差统计

维	幅度均值	期望权值	幅度方差	期望权值
1	497.87	1.00	217.65	1.00
2	239.61	2.08	139.76	1.56
3	274.36	1.81	154.46	1.41
4	165.68	3.01	128.95	1.69
5	89.87	5.54	70.98	3.07
6	80.37	6.20	57.02	3.82
7	101.31	4.91	62.12	3.50
8	56.71	8.78	39.65	5.49
9	69.86	7.13	46.47	4.68
10	43.50	11.45	31.99	6.80
11	38.48	12.94	30.76	7.07
12	39.24	12.69	30.78	7.07
13	31.48	15.82	24.98	8.71
14	26.84	18.55	22.92	9.50
15	24.58	20.26	20.46	10.64
16	22.86	21.78	18.72	11.63

有了这样的统计结果，同时兼顾计算的简单性，我们把权矢量取为 W=(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16).

§5.1.2　VQ 码本的确定

VQ 码本的优劣直接影响 VQ 的质量和系统的识别率，为了选择一个比较好的码本，我们分别用 LBG 算法和模拟退火 K-均值(ALK-means)算法对 97,615 个 CEP 矢量进行聚类。两种方法所用时间和所聚码本的平均失真度如表 5.2 所示：

表 5.2　两种聚类算法的比较

聚类算法	所用时间	平均失真
LBG	7h 52m 33s	1136.235433
ALK	2h 19m 53s	1232.784888

其中 ALK-均值的初始温度为 4℃。我们发现，LBG 算法虽然用的时间长，但码本的平均失真却经较小，因此系统选用由它产生的码本。至于 ALK-均值算法为什么没能取得比较好的效果，是因为初值选取不当还是其它原因，由于时间关系，我们没再作更多的实验。

§5.1.3 模型状态数的确定

模型的状态数选多少才算合适？为了弄清这个问题，我们用词表 2 分别取状态数为 3～6 做了实验，实验结果如表 5.3 所示。对男性或女性模型来说，错误率随模型状态数的增加而减小；而对综合模型来说，错误率在状态数为 5 时最小。由此可见，对非特定人数字识别系统来说，状态数取 5 是比较合适的！

表 5.3 模型状态数对识别率的影响

模型 \ 状态数 / 错误率	3	4	5	6
男性	2.846154%	2.538462%	2.307692%	2.230769%
女性	4.000000%	3.384615%	3.076923%	2.692308%
综合	6.423077%	5.500000%	4.769231%	4.884615%

§5.1.4 关于 VQ 的一些实验

为了弄清矢量量化(VQ)与非线性分段(NLP)的先后次序以及量化时是否使用码字矢量的类半径进行距离规整对系统识别率的影响，我们对词表 2 进行了实验，其结果如下表所示(模型状态数为 5)。从表中得出的结果我们发现：总的来说，量化时使用码字矢量的类半径进行距离规整(以下简称 rVQ)效果要好，而先进行 NLP 再 VQ 比先 VQ 再 NLP 效果绝对要好。

表 5.4 矢量量化与非线性分段的先后次序对识别率的影响

模型 \ 方式 / 错误率	VQ+NLP	NLP+VQ	rVQ+NLP	NLP+rVQ
男性	2.923077%	2.153846%	2.538462%	2.307692%
女性	3.769230%	2.461538%	3.615385%	2.000000%
综合	6.115385%	3.884615%	6.269231%	3.846154%

为了验证“rVQ”的效果，我们对词表 1 也做了实验，其结果如表 5.5(先 NLP 后 VQ)。结果是相似的：虽然“rVQ”在某些情况下效果较好，但总体效果并不令人满意。

表 5.5 rVQ 的效果(NLP+VQ)

模型 \ 方式 / 错误率	VQ	rVQ
男性	1.2%	1.2%
女性	3.0%	2.0%
综合	2.5%	3.1%

§5.1.5 关于识别时段长度规整对识别率影响的实验

在第三章，我们讲到用分段时段的长度对该段的计分值进行规整的公式：

$$F^{(w)}=\prod_{i=1}^{M}\left[\prod_{\substack{m_{i-1}<t\le m_i\\ C_t=V_j}} b_{ij}^{(w)}\right]^{1/L_i}$$

这个想法是基于理论分析与实验验证的。表 5.6 是对词表 1 进行实验的结果，其中模型状态数取 5。

表 5.6 段长度规整对识别率的影响

模型 \ 方式 / 错误率	VQ	rVQ	段长度规整 VQ	段长度规整 rVQ
男性	1.2%	1.2%	1.2%	0.8%
女性	3.0%	2.0%	2.0%	2.2%
综合	2.5%	3.1%	1.8%	2.7%

§5.1.6 有关声调评价的实验

我们对实验用的 100 遍语音数据中的‘0’~‘9’进行了声调检测，声调评价错误率 15‰。所谓声调评价，就是对某一数字的发音分别给出它是一声类、四声类和二三声类的要率。因此，声调评价错误是指三类中概率最大的类标签与实际的不符合，比如，我们对数字‘5’进行声调评价得出三类的概率分别为 0.8，0.6 和 0.7。从下面的实验结果可以看出，即使发生声调评价错误，总体识别率并不下降。表 5.7 是对词表 1(状态数 5，使用段长度规整)进行实验的结果。

表 5.7 声调评价对识别率的影响

模型 \ 方式 / 错误率	VQ	rVQ	声调评价 VQ	声调评价 rVQ
男性	1.2%	1.2%	0.0%	0.0%
女性	3.0%	2.0%	0.6%	1.0%
综合	2.5%	3.1%	0.7%	1.0%

§5.2 系统框图

经过上述实验，我们建立了一个系统。系统使用状态数为 5 的分段概率统计模型进行语音识别；在进行 VQ 时并不使用码字类半径进行距离规整；考虑到同一数字处在数字串不同位置或受与之相邻的音的影响，其发音会有所不同，系统共设四套模型或称模板；在进行识别时，各段的计分值用段的长度进行规整。四套模型中的每一套模型都有‘0’~‘9’十个模板，四套模型分别对应数字在串的

“首部”、“中部”、“尾部”和“单独”发音时的四种情形。模型的训练数据来自二十位男性：有十人每人发音两遍，每一遍包括一组单音数字‘0’~‘9’和一组三数字串(取自‘000’~‘099’，…，‘900’~‘999’十组中的一组)，另十个各发五遍单音数字‘0’~‘9’。图 5.1 是系统的框图。

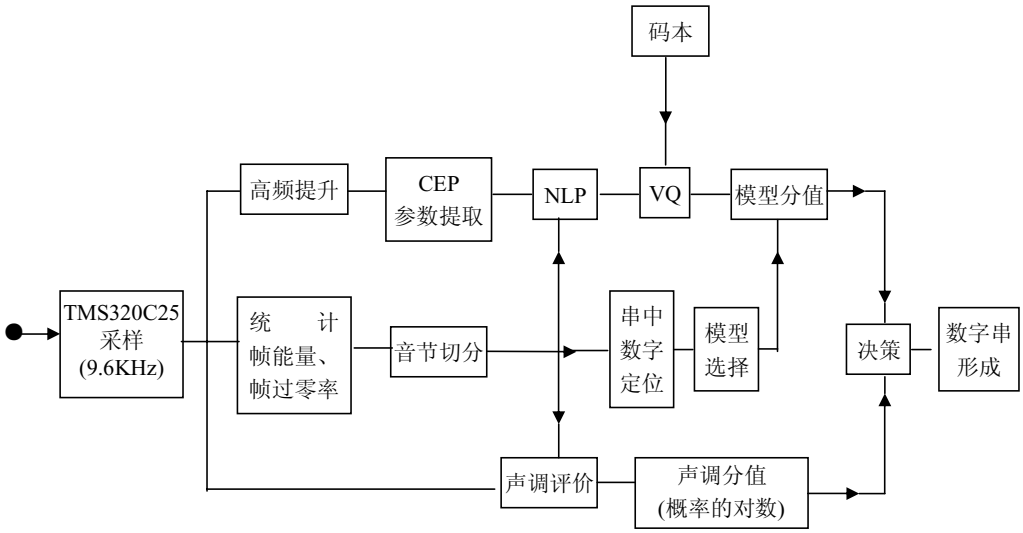

图 5.1 非特定人连续数字识别系统框图

§5.3 硬件配置

该系统由快速信号处理板(FSP)、IBM-PC 兼容机、一个话筒组成。

FSP 以 TMS320C25 为数字信号处理运算核心、微机为控制系统。TMS320C25 是 TMS320 数字信号处理器(DSP)系列中的一种 VLSI 数字信号处理器。TMS320 系列支持实时的数字信号处理，并广泛用于运程通讯、MODEM、语音处理、图象处理、频谱分析等领域。

TMS320C25 的总体设计强调其总的速度、通信及处理器结构的灵活性。控制信号和指令提供了块存储器传送、与片外慢速设备通信和多重处理的实现。通过单指令周期的乘/累加指令、两个片内大 RAM 块、八个辅助寄存器和专用运算部件、一个串行口、硬件定时器、用于数据扩充信号处理的快速 I/O 和许多别的特点大大增加了数字信号处理(DSP)应用的吞吐量。

FSP 系统资源分为两部分：

(一) TMS320C25 所具有的资源

64KW 程序存储器 PM
64KW 数据存储器 DM
2KW 双口数据存储器 DPDM
14 位带有可编程防折叠滤波器的 A/D 转换器
14 位带有可编程重构滤波器的 D/A 转换器
16 位并行输入口
16 位并行输出口
与主机间的 8 位/16 位方式的数据交换电路以及中断申请/中断复位电

路。

扩展口。

(二) FSP 所占微机的资源

I/O 地址空间(端口)：

① 0F20h：微型机用于向 TMS320C25 发控制字

② 0F21h：微型机用于向 TMS320C25 发中断申请

③ 0F22h：微型机用于清除 TMS320C25 发来的中断

存储器地址空间：用户根据所用微机的具体情况，通过调整 FSP 中的开关可决定占用微型机的下列五段之一：A 段(A000h 段)、B 段(B000h 段)、C 段(C000h 段)、D 段(D000h 段)以及 E 段(E000h 段)。

TMS320C25 提供了三个独立的地址空间，DM、PM 和 I/O 空间。FSP 开发板上的 PM 和 DM 由 TMS320C25 和微型机共同管理。片内的总裁逻辑电路保证 TMS320C25 和微机在同时执行程序时在 2K 的双口数据 RAM 中交换数据。而微机在管理 TMS320C25 的 260KB(128+128+4)的存储空间时，并不是把这些存储空间直接映射到微机中，而是采用窗口技术进行管理。窗口宽度为 64KB(微机中的一段)，即 FSP 所占用的微机的那一段地址空间。当微机要管理 FSP 的那部分存储器时，只需向控制端口发一个相应的控制字而把窗口对准相应部分即可。

TMS320C25 的控制由片内定时器、循环计数器、三种外部可屏蔽用户中断、串行口操作或定时器产生的内部中断提供的。

声频处理芯片 TLC32044，主要功能有 14 位的 A/D 和 14 位的 D/A 转换器以及可编程的防折叠滤波器和重构滤波器。TLC32044 在复位并启动后，利用内部中断 XINT 或 RINT，通过存取存储器映象寄存器 DXR 或 DRR 即可向 TLC32044 发送数据(D/A)或从 TLC32044 接收数据(A/D)。

§5.4 识别结果

我们随机产生了长度从 3～13 的数字串各 50 来个，并对该系统进行了无长度知识的数字串识别测试，得到了下面的实验结果，如表 5.8。其中“识别率 1”表示串识别率，“识别率 2”表示数字识别率。

表 5.8 系统识别性能测试(串识别率)

串长	3	4	5	6	7	8	9	10	11	12	13
识别率 1	96.0	94.0	90.6	88.5	87.8	86.0	84.0	82.0	78.0	78.0	74.0
识别率 2	98.7	97.5	96.2	98.1	98.0	98.3	98.0	98.0	98.0	98.0	98.0

第六章　汉语语音数据库的建立

目前尽管人们在汉语普通话的分析、识别和理解方面有很大兴趣，但还没有人在建立一个涉及大量的不同口音、不同年龄和不同性别的发音者的可理解汉语语音库方面做过系统的工作。

§6.1　建立语音数据库的意义与原则

虽在语音识别的算法应该是与语言种类无关，但事实上不可避免地有很多识别技巧是针对特定语种的，因此，世界各国都在动用大量的人力和物力建立适合自己语种的标准语音库，以适应语音识别理论研究和实践的不断深入。

建立标准的语音识别数据库，其意义主要反映在以下几点：

[1]提供实验分析数据：语音信号强烈依赖于发声者的个人特征，而且具有很大的随机性，为了进行全音节、孤立词、连接词和连接语音识别的研究，都需要在不同程度上对大量的实际语音数据来作统计分析(尤其是对非特定人的语音识别)，反复地用大量的语音样本来测试、修正识别算法和对识别系统进行训练，才能获得很好的识别性能。

[2]提供测试评价数据：标准的语音数据库可以为不同的语音识别算法和系统提供较为公正合理的测试评价数据，从而推动语音识别的发展。

[3]提供国际化共享资源：标准化的语音数据库更有利于国际上语音识别研究的交流，使得彼此都能共享这些资源，促进世界范围内语音识别研究的发展。

所以近十年来不少发达国家相继建立了本国语言的的标准语音数据库，并且这方面的国际合作和交流也正在形成。目前，国际上首先推出的语音数据库首当美国 DARPA 的含有 997 个词的海军资源管理数据库，美国 MIT、TI、SRI 等单位联合设计与制作、战略计算工程规划中应中 TIMIT 数据库[Lame1 86，Garofolo 88]，其次有日本 ATR(Advanced Telecommunications Research)研究所的语音数据库[Kurematsu 89]；英国的 SCRIBE 语音数据库，法国的 BDSON 数据库等等。

汉语是世界上使用人数最多的语种之一，汉字一字一个音节，400 多个无调音节和四声构成了丰富多彩的汉语词汇和语句。汉语不但有很多独特的个性特点，而且，即使讲普通话，不同地区的人也都带着浓厚的方言口音。为此，从语音识别的需要出发，很有必要建立一个包括不同地区(口音)、不同性别、不同年龄的发音人，具有自己特色而实用的汉语普通话语音数据库，这对于语音识别、语音分析、甚至语言理解方面的研究工作都将有很大帮助。汉语可理解语音库(A Comprehensive Chinese Speech Corpus)的建立，不仅在提供一个国际标准库、提供语言无关库等方面为语音识别的研究打开了方面之门，而且也使得语音识别的研究具有更大的效率。

我们认为，一个比较全面的数据库应具备如下条件：

[1]语料覆盖面宽：数据库中的语言材料应包含全部汉语音音节、不同音节连续发音时可能出现的协同发音变化，以及一些能满足连续语音识别需要的词组、短语、句了和短文。所设计的语音材料能包括语音特有的各种变化特征和群

体特征信息，满足各种不同目的的语音识别任务和需求。

[2]发音覆盖面宽：应有一定人数(约 200 到 500 名)的发音人，其中包括 2 小部分专业播音员，并考虑发音人年龄、性别和口音的适当分布。部分发音人对发音材料有多次发音。

[3]语音数据全面：数据库应不仅能提供数字化语音样本，而且能提供基本的声学特征参数，例如音素标记、声韵标记、词的分界标记、基音参数、LPC 参数、倒谱参数等等。

[4]数据库管理系统完善：数据模型的设计要使整个数据库所包含的数据冗余度降至最低，整个数据库结构合理，具有方便的数据录入、更新、检索和词汇编辑、安全维护等功能。有完善的输入输出接口，易于对语音识别系统进行训练、测试和评估。

§6.2 语音数据库介绍

我们组织了 235 个来自全国不同地区、年龄在 14～30 岁之间的男女学生及研究生进行了语音库的建立工作。语音库主要包括四个部分：

[1]单音节：单音节部分共有 20 组，每组有 76 个音节，所这些单音节字包括了汉语的 1334 个音节(附后)[Xian 85]。

[2]词组：词组部分分三组，它们是：

108 个二字词组

白桦	保守	被窝	便道	波动	不宜	惭愧	场合	衬衣
充满	储藏	磁化	打赌	大厦	当中	低级	电能	动力
短暂	发抖	犯人	匪军	风琴	负责	钢筋	根据	公债
瓜柄	规律	海棠	号码	宏伟	花费	荒谬	混纺	激动
继承	奸猾	酱油	揭发	今晚	精神	局面	开辟	可怜
酷似	狼狈	理由	粮站	溜脱	屡次	蔓延	迷藏	敏捷
木枪	内陆	努嘴	配合	平方	漆黑	恰巧	俏丽	清脆
圈子	人体	瑞雪	山垄	少数	审判	诗歌	使用	手劲
数据	思想	随便	滔滔	天然	挺拔	图画	外流	威武
温度	五星	细菌	现在	像片	斜阳	形状	许久	压制
燕山	野兽	依照	阴沉	用途	幼稚	圆晕	在于	战袍
针对	枝丫	纸伞	重量	住院	茁壮	总统	作战	蟋蟀

99 个三字词组

八达岭	白桦树	白茫茫	百花丛	百花园	班主任	办公室
北极星	博物馆	裁判员	长颈鹿	打火机	大理石	大气层
代表团	电视机	电影票	发明家	钢琴曲	高粱秸	工程师
光秃秃	红彤彤	花骨朵	吹呼声	黄花菜	会议厅	火车站
机械师	急行军	纪念馆	计算机	加拿大	建筑物	降落伞
教研组	金字塔	进行曲	井岗山	静悄悄	俱乐部	劳动力
老百姓	亮晶晶	林荫道	零用钱	旅游车	慢吞吞	毛茸茸
美滋滋	莫斯科	乒乓球	气象台	潜水艇	热水瓶	少年宫
圣诞节	实验室	手术台	数学题	水彩画	水电站	饲养员
探照灯	天安门	威尼斯	温度计	无线电	现代化	兴冲冲

摇钱树	野战军	意大利	萤火虫	邮递员	玉兰花	圆珠笔
运动会	照像机	自行车	作业本	正方形	冲锋枪	大会堂
大提琴	储水量	导火线	独木船	高粱米	黄鼠狼	教科书
龙王庙	南美洲	汽笛声	傻乎乎	体育场	心脏病	一瞬间
指南针						

102 个四字词组

白驹过隙	半信半疑	奔走相告	别无长物	不逞之徒
不堪设想	不期而会	不厌其详	残山剩水	豺狼当道
瞠目结舌	赤子之心	除恶务尽	蠢蠢欲动	大打出手
大显神通	倒海翻江	地旷人稀	洞烛其奸	咄咄怪事
反复无常	费尽心机	风雨交加	俯首听命	感同身受
根深柢固	孤苦伶仃	官逼民反	过甚其词	好高骛远
回光返照	祸起萧墙	疾首蹙额	艰难竭蹶	娇生惯养
解衣推食	尽如人意	敬而远之	据理力争	刻守不渝
滥竽充数	冷眼旁观	连篇累牍	灵丹妙药	屡次三番
满园春风	面黄肌瘦	明哲保身	目不识丁	匿影藏行
判若去泥	平白无故	欺人太甚	弃甲曳兵	前因后果
青黄不接	穷年累月	犬马之劳	人之常情	如胶似漆
弱肉强食	山高水低	舍死忘生	殊深轸念	铄石流金
随心所欲	体贴入微	停滞不前	投畀豺虎	唾手可得
万紫千红	为非作歹	畏葸不前	无出其右	无所适从
物以类聚	先声夺人	逍遥法外	心劳日拙	星星之火
休戚相关	血口喷人	言简意赅	养尊处优	一尘不染
一了百了	一手遮天	一张一弛	以已度人	易如反掌
饮鸩止渴	油头滑脑	鱼贯而行	源源不断	责有攸归
朝不谋夕	郑重其事	栉风沐雨	珠圆玉润	自生自灭
左顾右盼	坐享其成			

[3]短语：短语部分分三组，它们是：

Group 1 of 16 digit Strings

820599	608849	0701789	255739
687792	248195	644238	531358
618509	86722	152741	404329
510375	471211	3345	830028

Group 2 of 11 Multi-syllable Words

偶函数	恩赐	耳背	马奶
日记	承担	克服	绿叶
土尔其	破袜子	丝竹管弦	

Group 3 of 3 Confusable Uterances

傻娃娃怕他阿爸
义弟替你洗荸荠
祖父姑母数葫芦

[4]句子：句子部分共有 200 组，每组有 10 个句子，这些句子取自科技、科

普类报刊和文章。

发音者每遍发音包括一组单音节、一组词组、三组短语和一组句子，发音以普通话为基础。

不同类的语音分别存在不同的文件中，称为一个语音库文件。每个语音库文件都有相同的文件结构，即“文件头”并跟若干个“记录”。用 MSC6.0 描述的“文件头”的结构如下：

```
typedef union {        /* 语音库文件头结构/联合        */
    struct {
        char    chassflag;          /* 类标志        */
        char    speakerid[16]       /* 发音人姓名    */
        char    age;                /* 发音人年龄    */
        char    accent;             /* 发音人口音    */
        char    sex;                /* 发音人性别    */
        char    groupid;            /* 发音组号      */
    };
    char        _reserv[32];        /*字节对齐       */
}SL_header;
```

其中类标志有四个：FFh(单音节库)、Feh(词组库)、FDh(短语库)和 FCh(句子库)；发音人口音用发音人家乡邮政编码的前两位表示；发音人姓别用’M’表示男、’F’表示女。“记录”的结构用 MSC6.0 描述如下：

```
typedef struct {                /* 语音库记录结构/联合          */
    union {
        struct {
            char validflag [2];          /* 55Aah 表示记录有效   */
            char CC_no;                  /* 汉字的个数           */
            char Sample_no[4]            /* 样本长度             */
        };
        char    _reserv [16]             /* 16 字节对齐          */
    };
    char        Chinese [64];            /* 汉字串(机内码)       */
    char        Syllable [64]            /* 汉字编码串           */
    int         SpeechWave [];           /* 数字化语音采样值     */
} SL_record;
```

其中 SpeechWave[]是用一组整型数(2 字节)表示的数字化语音采样值，其长度由*(long*)Sample_no 给出。汉字编码是该汉字在汉字集中的序号。汉字集中共有 7,740 个汉字，包括数字‘0’~‘9’(‘1’读 yao)。在汉字集中，多音节被认为是不同的汉字，它们排在汉字集中相邻的位置，有不同的编码。汉字集中的汉字包括了几乎所有的一二级汉字(附后)。

由于有些语料是要重复婵 的，因此同一组语料有几遍发音，而语音库文件

的名字也就由“类”、“组号”和“遍号”唯一确定。

在我们的语音库中并没有存其它语音特征参数，因为这些参数可以由语音库管理软件给出。语音库管理软件有下面的功能：

方便的语音录入、放音、查阅等功能。

可给出某单音汉字注音或多音汉字的读音候选。

可把指定的语音库文件的指定记录调入操作缓冲区。

可用简单整系数滤波器[Zong 88]对缓冲区中的语音数据进行低通滤波(§2.1.4)、高通滤波和带通滤波处理[Zong 88]，滤波器参数可选。

可对缓冲区中的语音数据进行基音检测、参数(短时能量、对零率、LPC、CEP 参数等)提取等操作。

可对缓冲区中的语音数据进行音节切分、音素切分等处理。

这部分工作由作者及李建民博士完成。

有了这个语音库，对我们的语音合成及识别的研究将是一个很大的促进。

参考文献

【Atal 68】B.S. Atal, M. R. Schroeder
Predictive coding of speech signals
Proc. 6th Int. Cong. Acoust; C-5-4, 1968

【Baker 75】J. K. Baker
The DRAGON System -An Overview
IEEE Trans. On ASSP-23(1), Feb; 1975

【Bakis 76】R. Rakis
Continuous Speech Recognition via Centisecond Acoustic States
In 91st Meeting of the Acoust. Soc. of Am; Apr; 1976

【Bahl 80】L.R Bahl, F. Jelinek, R.L. Mercer
A Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition
In Proc. of IEEE ICASSP-86, Apr; 1986

【Bahl 86】L. R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mereer
Maximum Mutual Informatiion Estimation of Hidden Markov Model Parameters for Speech Recognition
In Proe. of IEEE ICASSP-86, Apr; 1986

【Bahl 88a】L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer
A New Algorithm for the Estimation of Hidden Markov Model Parameters

【Bahl 88b】L. R. Bahl, P. E. Brown, P. V. de Souza, R. L. Mercer
Speech Recognition with Continuous Parameter Hidden Markov Models
In. Proc. Of IEEE ICASSP-88, PP40-3, Apr; 1988

【Baum 72】L. E. Baum
An Inequality and Associated Maximization Technique in Statisical Estimation of Probabilistic Functions of Markov Processes, Inequalities, 3, 1972

【Bourlard 89】H. Bourlard, C. J. Wellekens
Speech Dynamic and Recurrent Neural Networks
ICASSP-89, 1989

【Bride 82】J. S. Bride, et al
An Algorithm for Speech Recognition, ICASSP-82, 1982

【Brown 87】P. F. Brown
The Acoustic-Modeling Problem in Automatic Speech Recognition
PH. D. Thesis, Carnegie Mellon University, May, 1987

【Cao 89】曹洪
非特定人汉语孤立字识别系统研究
申请博士学位论文详细摘要，清华大学，1989

【Cravero 84】M. Cravero, L. Fissore, R. Pieraeeini, C. Scagliola
Syntax Driven Recognition of Connected Words by Markov Models,

In Proe. of IEEE ICASSP-84, Apr; 1984

【Davis 80】S. B. Davis, P. Mermelstein
Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences IEEE Trans. on ASSP-28(4), Aug; 1980

【DeMori 86】R. DeMori, L. Lam
Plan Refinement in a Knowledge-Based System for Automatic Speech Recognition, ICASSP-96, 1986

【Franzini 90】M. A. Franzini, K. F. Lee
Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition, ICASSP-90, 1990

【Furui 86】S. Furui
Speaker-Independent Isolated Word Recognition
Using Dynamic Features of Speech Spectrum
IEEE Trans. on ASSP-34(1), Feb; 1986

【Furui 88】S. Furui
A VQ-Based Preprocessor Using Cepstral Dynamic
Features for Speaker-Independent Large
Vocabulary Word Recognition
IEEE Trans. on ASSP-36(7), Jul; 1988

【Furui 89】Sadaoki Furui
Digtal Speech Processing, Synthesis, and Recognition
MARCEL DEKKER, INC; 1989

【Gao 88】高雨青，陈永彬，吴伯修
汉语非特定人连续数字识别的一种实现
D6—8，CCSP—88

【Gao 90】Y. Gao, T. Huang, D. Chen
HMM-Based Warping in Neural Networks
ICASSP-90, 1990

【Garofolo 88】J. S. Garofolo
The Structure and Format of the DARPA TIMIT
CD-ROM Protype
Documentation of DABPA TIMIT, 1988

【Haton 84】J. P. Haton
Knowledge-Based and Expert Systems in Automatic
Speech Recognition, in R. DeMori (ed), New Systems and Architectures for Automatic Speech Recognition and Synthesis, 1984

【Huang 89】X. D. Huang, M. A. Jack
Semi-Continuous Hidden Markov Models for Speech
Signals, Comeputer Speech and Language No.3, 1989

【Itakura 68】F. Itakura, S. Saito
Analysis synthesis telephony based on maximum like-lihood method, Proc. 6th Int. Cong. Acoust; C-5-5, 1968

【Jelink 76】F. Jelink
Continuous Speech Recognition by Statistical Methods

Proceedings of the IEEE, 64(4), Apr; 1976

【Jiang 89】蒋力
基于概率统计模型的非特定人语音识别方法与系统的研究清华大学计算机系硕士学位论文，1989.11

【Juang 85a】B. H. Juang, L. R. Rabiner
Recent Developments in the Application of Hidden Markov Models to Speaker-Independent Isolated
Word Recognition, In IEEE ICASSP-85, Apr; 1985

【Juang 85b】B.H. Juang, L. R. Rabiner
Mixture Autoregressive Hidden Markov Models for
Speech Signals, IEEE Trans. on ASSP-33(6), Dec; 1985

【Kirkpatric 83】S. Kirkpatric, D.D. Galatt, M.P. Vecchi
Optimization by Simulated Annealing, Sci; 220, 1983

【Klatt 80】D. H. Klatt
Overview of ARPA-Speech Understanding Project in New Trends in Speech Becognition, Prentice Hall, 1980

【Kohonen 84】T. Kohonen
Self-Organization and Association Memory
Springer-verlab, Berlin, 1984

【Kurematsu 89】A. Kurematsu, K. Takeda, H.Kuwabara, K. Shikano
ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis

【Lamel 86】L. F. Lamel, R. H. Kassel, S. Seneff
Speech Database Developmemt: Design and Analysis of the Acoustic- Phonetic Corpus
Proc. Speech Recognition Workshop, 1968 (DARPA)

【Lee 88a】K. F. Lee
Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System
PH. D. Thesis, Carnege-Mellon University, Apr; 1988

【Lee 88b】K.F. Lee, H. W. Hon
Large-Vocabulary Speaker-Independent Continuous Speech Recognition
In Proc. of IEEE ICASSP-88, PP123-6, Apr; 1988

【Levison 83】S.E. Levison, L.R. Rabiner, M.M. Sondhi
An Introduction to the Application of the Theory of Probabilistic Function of A Markov Process to Automatic Speech Recognition
Bell Syst. Tech. Journal, Vol 62(4), Apr; 1983

【Li 90】李建民
基于音节的大字表语音识别系统的研究和实现
清华大学计算机系硕士学位论文，1990.6

【Linde 80】Y. Linde, A. Buzo, R.M. Gray
An Algorithm for Vector Quantization Design
IEEE Trans. on COM-28(1), Jan; 1980

【Lu 86】陆大 金

随机过程及其应用
清华大学出版社，1986.8

【Markel 76】J.D. Markel, A.H. Gray
Linear Prediction of Speech
Springer-Verlag, New York, 1976

【Mao 88a】茅为华
用隐式马尔可夫模型构成非特定人语音识别系统的研究
清华大学计算机系硕士学位论文，1988.6

【Mao 88b】W. H. Mao, L. Jiang, W. H. Wu, D. T. Fang
An Introduction to a Demonstrating Recognition System For Speaker-Independent Isolated Chinese Speech
In Proc. of IEEE Int. Workshop on ATSA, Aug; 1988, Beijing

【Mao 88c】茅为华，蒋力，吴文虎
用隐式马尔可夫模型构成非特定人语音识别系统的研究
第二届全国人工智能与模式识别会议论文，1988.10

【MeDermott 89】E. McDermott, et al
Shift-Invariant, Multi-Category Phoneme Recognition
Using Kohonen's LVQ2, ICASSP-89, 1989

【Merialdo 88】B. Merialdo
Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information Training
In Proc. of IEEE ICASSP-88, Apr; 1988

【Morgan 90】N. Morgan, H. Bourlard
Continuous Speech Recognition Using Multilayer Perception with Hidden Markov Models
ICASSP-90, 1990

【Myers 81】C. S. Myers, L.R. Babiner
A Level-Building Dynamic Time Wraping Algorithm for connected Word Recognition
IEEE Trans on ASSP, Vol. 29, No. 2, April 1981

【Nadas 83】A. Nadas
A Decision Theoretic Formulation of a Training Problem in Speech Recognition and a Comparison of Training by Unconditional versus Conditional Maximum Likelihood
IEEE Trans. on ASSP-31(4), Aug; 1983

【Nadas 88】A. Nadas, D. Nahamoo, M.A. Picheny
On a Model-Robust Training Method for Speech Recog-nition, IEEE Trans. on ASSP-36(9), Sep; 1988

【Ney 88】H. Ney, A. Noll
Phoneme Modeling Using Continuous Mixture Densities
In Proc. of IEEE ICASSP-88, Apr; 1988

【Paul 86】D. B. Paul, R. P. Lippmann, Y. Chen, C. Weinstein
Robust HMM-Based Techniques for Recognition of Speech Produced under Stress and Noise
Speech Tech; Apr; 1986

【Paul 88】D. B. Paul, E. A. Martin
Speaker Stress-Resistant Continuous Speech Recognition
In Proc. of IEEE ICASSP-88, Apr; 1988

【Babiner 76】L. R. Rabiner, M.R. Sambur
Some Preliminary Experiments in the Recognition of Connected Digits
IEEE Trans. on ASSP, Vol.24, No.2, April 1976

【Babiner 78】L. R. Rabiner, R.W. Schafer
Digital Precessing of Speech Signals, Prentice-Hall, 1978

【Rabiner 81】L. R. Rabiner, J. G. Wlipon
Isolated Word Recognition Using a Two-Pass Pattern Recognition Approach, ICASSP-81, Mar; 1981

【Rabiner 83】L. R. Rabiner, S. E. Levison, M.M. Sondhi
On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition, Bell Syst. Tech. Journal, Vol 62(4), Apr; 1983

【Rabiner 84】L. R. Rabiner, K. C. Pan, F. K. Soong
On the Performance of Isolated Word Speech Recognizer Using Vector Quantization and Temporal Energy Contours
AT&T Bell Tech. Journal, 63(7), Sep; 1984

【Rabiner 85】L. R. Rabiner, J. G. Wilpon, B. H. Juang
A Segmental K-Means Training Procedure for Connected Word Recognition, AT&T Tech. Journal Vol 65(3), 1985

【Rabiner 86a】Rabiner L. R., Juang B. H., Levison S. E, Sondhi M.M.
Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities
AT&T Tech. Journal 64(6), Jul./Aug; 1986

【Rabiner 86b】L. R. Rabiner, B. H. Juang
An Introduction to Hidden Markov Models
IEEE ASSP Magazine, 3(1), Jan; 1986

【Rabiner 88】L. R. Rabiner, J. G. Wilpon, F. K. Soong
High Performance Connected Digit Recognition Using Hidden Markov Models
In Proc. of IEEE ICASSP-88, Apr; 1988

【Rabiner 89】L. R. Rabiner, J. G. Wlipon, F. K. Soong
High Performance Connected Digit Recognition Using Hidden Markov Models
IEEE Trans. on ASSP Vol.37, No.8, August 1989

【Russell 85】M. J. Russell, R. K. Moore
Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition
In Proc. of IEEE ICASSP-85, Apr; 1985

【Sakoe 79】H. Sakoe
Two Level DP Matching-A Dynamic Programming based Pattern Matching Algorithm for Connected Word Recognition

IEEE Trans. on ASSP Vol.27, No.6, Dec. 1979

【Sakoe 89】H. Sakoe, et al
Speaker-Independent Word Recognition Using Dynamic Progamming Neural Networks, ICASSP-89, 1989

【Schwartz 84】R. M. Schwartz, Y. L. Chow, S. Roucos, M. Krasner, J.Makhoul
Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition
In Proc. of IEEE ICASSP-84, Apr; 1984

【Shikano 86】K. Shikano, K. F. Lee, D. R. Reddy
Speaker Adaptation through Vector Quantization
In IEEE ICASSP-86, Apr; 1986

【Soong 86】F. K. Soong, A.E. Rosenberg
On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition
In Proc. of IEEE ICASSP-86, PP877-80, 1986

【Sugawara 85】K. Sugawara, M. Nishimura, K.Toshioka, M.Okochi, T.Kaneko
Isolated Word Recognition Using Hidden Markov Models
In Proc. of IEEE ICASSP-85, Apr; 1985

【Tebelskis 90】J. Tebelakis, A. Waibel
Large Vocabulary Recognition by Linked Predictive Neural Networks, ICASSP-90, 1990

【TI 88】Texas Instruments Corp.
TMS320C25 User's Guide

【TMS 89】TMS320C25-E 型开发/高速处理板使用说明
中国科学院声学所十室，1989

【Tohkura 86】Y. Tokura
A Weighted Cpstral Distance Measure for Speech Recognition, In IEEE ICASSP-86, Apr; 1986

【Vintsjuk 68】T. K Vintsjuk
Recognition of Words of Oral Speech by Dynamic Programming, Kibernetica, Vol. 81, No.8, 1968

【Viterbi 67】A.J. Viterb
Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm
IEEE Trans. on IT-13(2), Apr; 1967

【Waibel 88】A. Weibel, et al
Phoneme Recognition: Neural Networks vs Hidden Markov Models, ICASSP-88, 1988

【Wang 88】王作英，曹洪
语音识别的改进隐含马尔可夫模型
智能计算机学术会议论文，1988

【Wang 89】王作英
基于段长分布的隐含马尔可夫模型
全国第三届汉字与语音识别会议论文，1989

【Wiener 66】N. Wiener

Extrapolation Interpolation and smoothing of Stationary Time Series
MIT Press, Cambridge, Massachusetts, 1966

【Xian 85】现代汉语词典
中国社会科学院语言研究所词典编辑室编，商务印书馆，1985

【Xu 89】徐雷
一类新的聚类分析算法：模拟退火法
《模式识别与人工智能》第二卷第 1 期，1989.3.

【Yuan 90】苑宝生，俞铁城，应海芳
连呼汉语数字识别及其应用系统
中国科学院声学研究所，1990

【Zhao 90】赵彤青
汉语语音输入系统的研究
清华大学计算机系学士学位论文，1990.6.

【Zheng 90】郑方
语音端点检测、前端处理和特征抽取的研究
清华大学计算机系学士学位论文，1990.6.

【Zheng 92】郑方，吴文虎
汉语连续语音识别中音节自动切分的研究
第四届全国汉字及语音识别学术会议论文，1992.5.

【Zong 88】宗孔德，胡广书
数字信号处理
清华大学出版社，1988.6.

【Zue 85】V.W. Zue
The Use of Speech Knowledge in Automatic Speech Rcognition, Proceeding of the IEEE, 73(11), Nov; 1985

【Zue 89】V. W. Zue, et al
Acoustic Segmentation and Phonetic Classification in the SUMMIT System, ICASSP-89, 1989

【Zue 90】V. W. Zue, et al
The SUMMIT Speech Recognition System:
Phonological Modeling and Lexical Access
ICASSP-90, 1990

【Zwicder 61】E. Zwicker
Subdivision of the Audible Frequency Range into Critical Bands
Journal of Acoust. Soc. Am. 33, Feb; 1961

附录一：1334 个汉语拼音(略)

附录二：7740 个汉字(略)