第二届全国人机语音通讯会议论文集(NCMMSC-92)，pp.160-165，1992 年 9 月，桂林

非特定人连续汉语数字识别方法与系统

郑　方　吴文虎　方棣棠
清华大学计算机系，100084

摘要〗本文对非特定人连续数字识别方法进行了比较深入的研究。

连续数字的识别有着与其他语音识别不同的特点，数字串中各数字之间没有相关的知识，因此要求音节切分及数字识别的正确率都必须很高。为此，作者进行了大量的实验研究，确定使用基于非线性分块的分段概率模型作为识别模型，这是一种类似于隐马尔可夫模型的概率统计模型，但它却没有状态转移矩阵。通过实验，作者发现，当识别时计算参考模型各状态的分数估算时使用分段长度进行规整有很好的效果。作者还提出了把数字按声调分成三类的想法和数字的声调（概率）评价函数的概念，在给出某待识音节后，该函数给出它是三类中的任一类的概率值（三个），这三类是一声类、二三声类和四声类。由于声调知识并不用作粗分类，可避免因错误剪枝导致的识别率下降。

作者完成的非特定人连续数字识别系统基于段长度规整的分段概率模型，在识别时把待识音节对各参考模型的分值与该待识音节的声调（概率）评价函数值进行综合。同时，考虑到连续语音中毗邻音节间的相互影响，采用多套模型，它们分别对应于数字在串的“首部”、“中部”、“尾部”和“单独”发音时的四种情形。

使用上述的识别策略建立的系统，其识别效果有相当大的改善。对长度为　3~13 的随机长度数字串，单个数字识别率全部都在 98.00% 以上，音节切分的正确率高达 99.64%。

关键词：非特定人连续数字识别，非线性分块，分段概率模型，声调评价函数

一、引言

人类真正的自然的语言是连续语音，如何解决连续语音的识别问题是语音通信技术的主攻方向之一。人们最早对语音识别的研究局限在　“小字表”、“特定人”和“孤立词”的范围内，随着方法和技术的不断成熟，识别领域分别沿三个方向向“大字表”、“非特定人”和“连接词（连续语音）”进行了拓展，而这些方面的工作和技术也正是语音识别的难点。在这些问题尚未得到圆满解决之前，在其中一两个方面进行拓展对语音识别的研究无疑有较为广泛的实用价值。在这种情况下，本文对非特定人连续数字识别方法进行了研究，取得了令人鼓舞的效果。

二、原理与算法简介

1°端点检测和音节切分

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能量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，二要简便易行。

本系统根据帧平均幅度进行端点检测，用帧平均幅度、帧过零率及音长信息为判据，使用相对

（幅度）阈值进行音节切分[Zheng 92]，收到很好的效果，音节切分的正确率高达 99.64%。

•“帧平均幅度”是指一帧语音样值的平均幅度。其计算公式为：

$$FENG = \frac{1}{N}\sum_{i=0}^{N-1}|S_n(i)|$$

•“帧过零率”是指一帧语音的短时过零数。其计算公式为：

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

$$\delta_i = \begin{cases} 1, & |S_n(i)| \geq ZLEV \text{ 而且 } S_n(i) * S_n(i-1) < 0 \\ 0, & \text{其他} \end{cases}$$

在公式中，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，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。

本系统帧长为 256 点（9.6ＫＨｚ采样率），帧移为 128 点。

２° 特征参数

LPC-CEP（倒谱）系数是一种比较好的特征参数，　本系统选用１６阶 LPC-CEP 系数作为特征参数。在一帧语音进行参数提取前先进行高频提升：

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

在实验中我们发现，CEP 系数各维的幅度均值和方差并不均衡，为了使各维分量的贡献比较相当，我们对其各维进行加权，权向量是经过对大量 CEP 向量的各维的幅度均值和方差统计得出的。

３° 非线性分块原理

非线性分块（Non-Linear Partition）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。

设有语音观察序列 O＝｛C_1，…，C_T｝，其中 C_i 为 K＝16 阶 CEP 系数矢量，C_i＝($C_i(1)$，…，$C_i(K)$)，T 为观察序列长度。要将观察序列分成 M 块，O＝P_1+P_2+…+P_M（其中“+”表示串接），为此定义语音特征变化信息为：

$$d_j = dcep(C_j, C_{j+1}) = \sum_{k=1}^{K} W_k (C_j(k) - C_{j+1}(k))^2, \quad 1 \leq j \leq T-1$$

其中 W＝(W_1，…，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($1 \leqslant i \leqslant M$，令 m_0＝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。

4° 分段概率模型原理

无论从理论还是从实践来看，隐马尔可夫模型（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。但对于模型的结构究竟是否符合人类语音产生的机理以及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），取得了较好的效果。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。蒋　力提出了一种基于 NLP 的分段概率模型（Segmental Probabilistic 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。作者在此基础上，对算法作了一些改动，使得识别率进一步提高。

下面对这一原理作一简单的介绍。设模型状态数（即分块数）为 M，模型的状态为

$S=\{S_i\}$，　$1 \leqslant i \leqslant M$

码本长度为 NC，码本为

$V=\{V_i\}$，　$1 \leqslant i \leqslant NC$

模型的输出概率矩阵为

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

假设训练遍数为 L，则 B 矩阵的计算公式为：

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

显然 $\sum_j b_{ij} = 1$。

识别时求出待识字与每个模型字 W_k 匹配的计分值：

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

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

这种模型与 HMM 相比，仅有输出概率信息而无状态转移信息。它虽是一个概率统计模型，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：

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

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。

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，给 B 矩阵的零元素赋予一个最小值ε [Levinson 83]，然后修改 B 矩阵的其他元素以满足约束条件。

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

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

一般认为ε值应选在 10^{-6}～10^{-4} 之间[Levinson　83],在我们的系统中选取ε＝10^{-4}。

SPM 的状态数 M（即非线性分块的块数）对识别率会有什么影响呢？我们经过实验后发现，在 M＜ 5 时识别率随 M 的增大而增大，但 M> 5 后效果并不明显，因此我们认为状态数取 M=5 对数字识别来说是比较合适的。

考虑到同一个数字处在数字串的不同位置或受与之相邻的音的影响，其发音会有所不同，我们使用四套模型，它们分别是数字在串的“头部”、“中部”、“尾部”及“单独”发音时训练得到的。这种作法使识别率进一步提高。

5 ° 实际考虑

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），从而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

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

这样做可大大简化计算。

6 ° 声调评价

对于一个数字识别系统，可以把它分为三类：一声类（数字 1、3、7、8）、二三声类（数字 0、5、9）和四声类（数字 2、4、6）。实验表明，这种分法具有类内可分性。

在对某一音节检测出基音周期估值序列｛P_i｝后，从中抽出两点（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。

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 1 可以看出，这样做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。

使用上述方法判断出三类声调后，将其结果以概率的形式参加总体的识别，而不是简单地用于粗分类，这样做防止了由于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 5 9 5 发成为　wú　jiu　wǔ ），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。

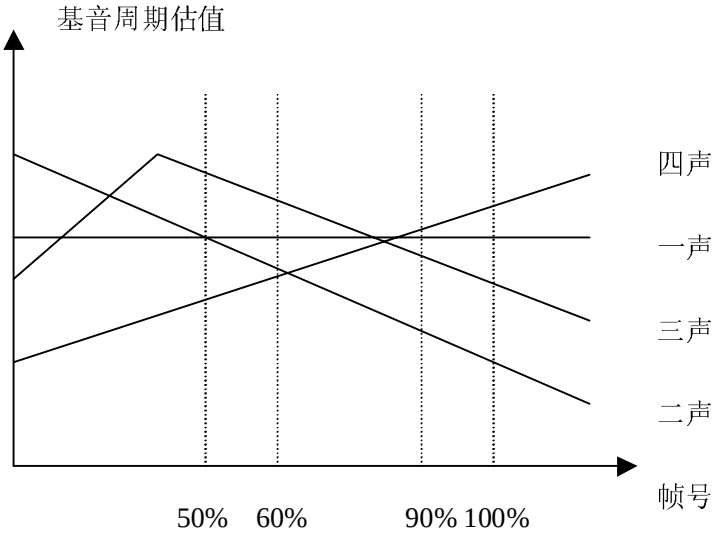

图 1　汉语四个声调的基音周期走向图

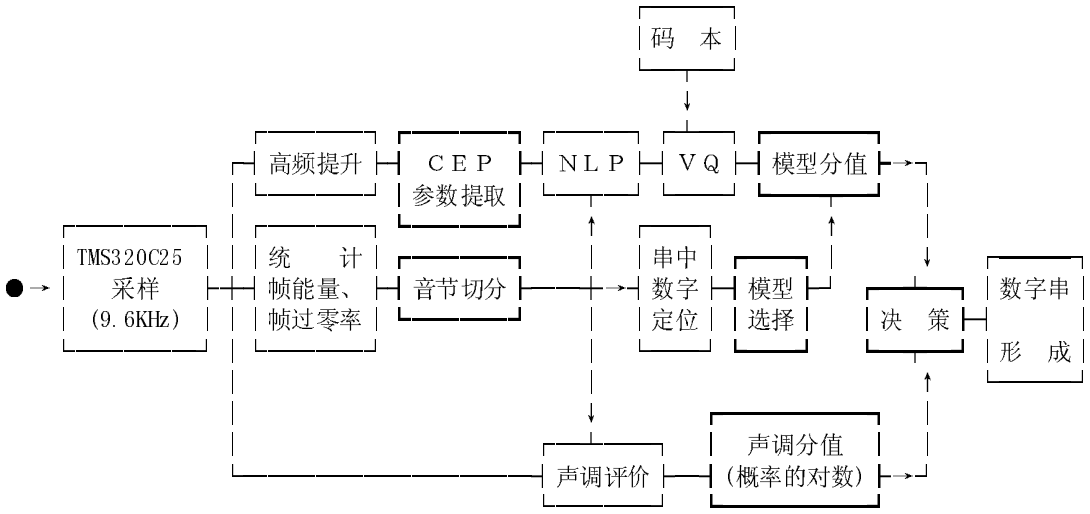

图 2　系统框图

三、系统构成

图 2 是系统的框图。我们随机产生了长度为 3 ～ 1 3 的数字串若干个，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 1。在统计时，只要数字串的音节切分错误或数字串中有一个数字识别错误就认为这个“串”误识。

表 1 系统识别率测试结果

串长	3	4	5	6	7	8	9	10	11	12	13
测试串个数	50	50	53	52	49	50	50	50	50	50	50
数字串误识	2	3	5	6	6	7	8	9	11	11	13
串识别率	96.0	94.0	90.6	88.5	87.8	86.0	84.0	82.0	78.0	78.0	74.0
数字误识	2	3	5	6	7	7	9	9	11	12	13
数字识别率	98.7	98.5	98.1	98.1	98.0	98.3	98.0	98.2	98.0	98.0	98.0

【主要参考书】

〖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

〖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

〖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

〖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”，1992.5