第 27 卷 第 5 期　　物 理 学 报　　Vol. 27, No. 5
1978 年 9 月　　ACTA PHYSICA SINICA　　Sept., 1978

通用实时语言识别系统——RTSRS(01)*

俞 铁 城
(中国科学院物理研究所)

提　　要

本文描述一个通用实时语言识别系统——RTSRS(01). 在以前工作的基础上[1],每条口呼命令的参数在时间域上规正, 采用二值频谱[2,3],大大压缩了参考音的参数存贮量, 同时应用新的求差距的办法,使得识别所需的时间大为缩短,以致字表为 200 时能实时识别. 专人的识别结果为: 口呼数字, 99.7%; 20 句话(每句 7 字), 99.7%; 四字成语 100 个, 99.5%; 四字成语 150 个, 99.3%; 四字成语 200 个, 98.8%; 四字成语 400 个, 99.7%. 非正式的实验表明,对于不同音节数的字表,乃至口呼英语数字或 BASIC 语句名字等,都有高的正确识别率.

一、引　　言

口呼命令的机器识别, 对于自动控制、语言通信、计算机的口呼输入等等具有现实的重要意义. 我们以前的工作[1],做到了字表大小为 20 的情况下,专人识别率达 99%以上.

要在小型计算机(8K 内存,无外存)上能识别较大规模的字表,并使尽量实现实时识别,必须对方案作进一步的革新. 我们采用二值频谱作为参数. 同时,对于差距的计算也相应地采用新的方法,经过一段时期的工作,取得了良好的结果.

二、选谱方案的改进

以前的方案[1]中,从频谱的声刺激概念出发, 对每一条口呼命令在时间域上非线性地规正发音长短,选出一定数目的频谱构成识别图样. 现在作了两点改进:

1. 允许重复选中某个频谱

从以前的工作过程中,注意到音变化的时候, 频谱声刺激量[1]往往很大. 图 1 和图 2 是几个口呼数字的频谱声刺激随时间变化的曲线图.这可能对于人的听觉有较大的贡献. 为模仿此点,在现在的方案中,允许重复选中某个频谱.

2. 撤消最后一个频谱必被选中的权利

以前的方案中,最后一个频谱必定选中, 这在单音节语音, 选用频谱数目很少的情况下,很不理想,也不符合预定的"随机选定"的原则,因此新方案中撤消这个必然性. 具体做法如下:

* 1977 年 11 月 30 日收到.

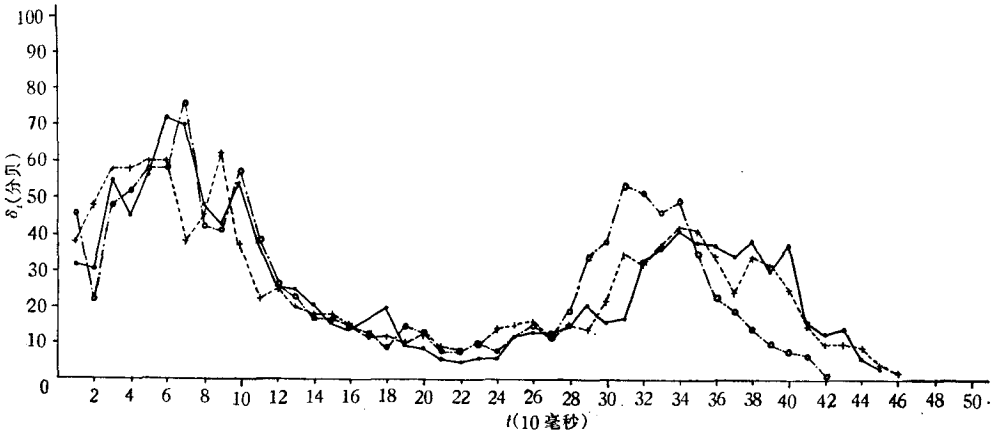

图 1　呼音(1)的频谱声刺激图（三次发音）
1976 年 1 月 14 日和 15 日数据

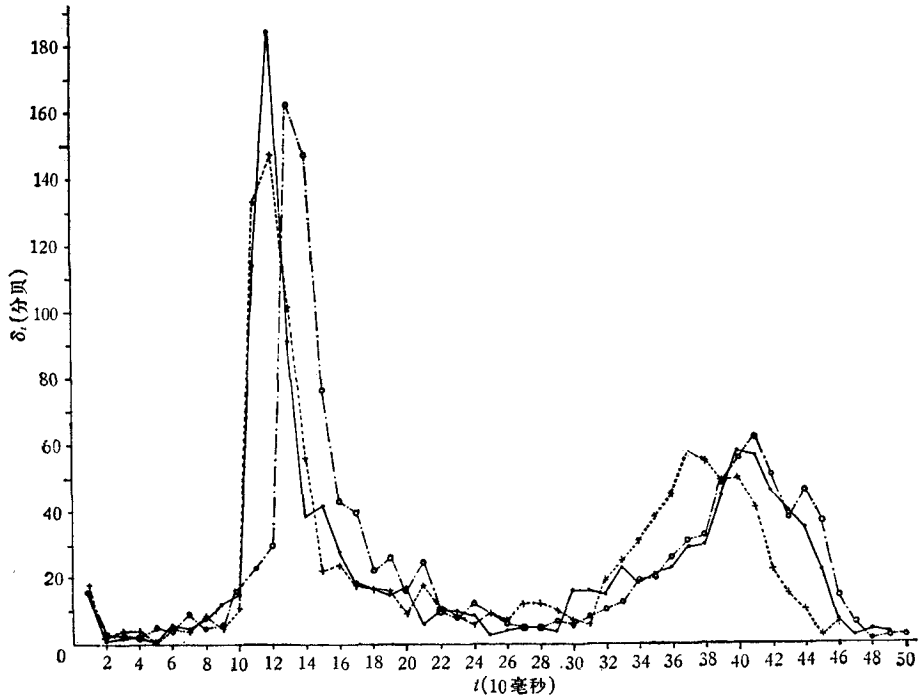

图 2　呼音(3)的频谱声刺激图(三次发音)
1976 年 1 月 14 日和 15 日数据

由文献[1],我们有相邻两时刻的频谱产生的声刺激为

$$\delta_1 = \sum_{j=1}^{L} A_{1j} \qquad n = 1;$$

$$\delta_n = \sum_{j=1}^{L} |A_{nj} - A_{n-1j}| \quad n = 2, 3, \cdots, N. \tag{1}$$

量 $|A_{nj} - A_{n-1j}|$ 称为第 j 频带里在时刻 n 产生的声刺激. 这条口呼命令产生的总声刺激便是

$$\Delta = \sum_{n=1}^{N} \delta_n. \tag{2}$$

如果要选出M个频谱来标征这条口呼命令,则找出所谓的频谱声刺激的"小"平均值为

$$\Delta_0 = \Delta/(M + 1). \tag{3}$$

寻找这M个频谱的方法是:

(1) 声刺激累加器清零 $R = 0$.

(2) 依次取下一个 δ_n($n = 1$ 是始值),加进声刺激累加器:

$$R \Leftarrow R + \delta_n.$$

(3) 若 $R \geqslant \Delta_0$, 则选中这个第 n 谱,转往(5).

(4) 否则,不选中这个第 n 谱,声刺激累加器只是增加了一个 δ_n, 转往(2).

(5) 将这个第 n 谱,经二值化处理后记住, 并给它编上选中谱序号 p. 同时, 让声刺激累加器 R 减少 Δ_0,

$$R \Leftarrow R - \Delta_0.$$

(6) 查一下,一共选中了M个频谱没有. 若够了,则告结束;否则转往(3)继续寻找.

在计算机内存允许及实时处理能力还具备的情况下,M宜选得稍大一点为好. 当然,只要不影响识别率,M越小越好,因为M小,内存就省,识别又快. 在我们作的一系列实验中, 当M选成 $M = (2—4) * S(\max)$, 其中 $S(\max)$ 为命令表中最长的命令里包含的音节数目,此时识别率不降低.

三、频谱的二值化处理

以前有人应用二值频谱来识别十个口呼数字[2],或几十个口呼字[3]. 现在我们把这种处理方法应用到字表较大的情况. 二值化处理时所用的阈值之选取有所不同, 现说明如下:

假定被选中的第 p 个频谱原始数据(已是分贝值)是

$$A_{p1}, A_{p2}, \cdots, A_{pL},$$

其中,右脚标 1, 2, …, L 指滤波通道序号. 在我们的 1/3 倍频程频谱分析系统中[1],谱动态范围是 50 分贝,低限对各通道是一样的,上述 A_{pj}, $j = 1, 2, \cdots, L$, 是高出这个低限的值(分贝值),所有低于或等于这个低限的数据均为零.所以,在L个 A_{pj} 中,往往有好几个是零. 为了克服音量变化时这零的个数的影响, 我们用非零数据的平均值作为二值化处理时的阈值. 今假定其中有 s 个值不为零,可用下式求出频谱的所谓"大"平均值:

$$\bar{A}_p = \sum_{j=1}^{L} A_{pj}/s. \tag{4}$$

然后按下式进行二值化处理:

$$a_{pj}=\begin{cases}1 & \text{当 } A_{pj}\geqslant \bar{A}_p \text{ 时;}\\ 0 & \text{当 } A_{pj}<\bar{A}_p \text{ 时.}\end{cases}\quad (j=1,2,\cdots,L) \tag{5}$$

对于每一个选中频谱都进行这样的二值化处理.（实际上，每一个原始频谱都进行二值化处理后存放的，因为在两次采样间隙约有 6 毫秒的等待时间，完全可以用来进行二值化处理，以便发完音之后很快得到识别结果）. 因此，二值化处理所用的阈值随频谱而变，它是浮动的而不是固定的[2]，这有助于在一定范围内克服音量变化的问题. 对于每一个选中频谱，我们只用一个阈值而不采用多个阈值[3]，运算简单，易于制备硬件.

经过二值化处理后，信息存贮量就大大缩小了. 现在，一个频谱分量只要 1 比特，而以前用浮点数表示时[1]，在我们的计算机（Varian 620/L）上要用 32 比特，这样就能节省内存三十二倍.

四、差距的求法

由于用二值频谱作为最后识别用的数据，就必须采用新的办法来计算被比较的两个音之间的距离.

假定一个频谱用 L 个通道数据来表示（当 $L=16$ 时，刚好占用一个 16 位字长的计算机字）. 将两个二值频谱进行比较时，只需进行一次异或（模 2 加），然后点一下有几个位上是 1，这 1 的个数就算作这两个二值频谱之间的距离. 假定每一条口呼命令选用 M 个二值频谱，字表大小为 B，则第 k 个参考音数据为

$$[SBP]^{(k)}=\begin{bmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & a_{1L}^{(k)}\\ a_{21}^{(k)} & a_{22}^{(k)} & \cdots & a_{2L}^{(k)}\\ \vdots & \vdots & & \vdots\\ a_{M1}^{(k)} & a_{M2}^{(k)} & \cdots & a_{ML}^{(k)}\end{bmatrix}\qquad k=1,2,\cdots,B. \tag{6}$$

其中，$a_{ij}^{(k)}=0$ 或 1，$i=1,2,\cdots,M$；$j=1,2,\cdots,L$.

新呼的待识别命令的二值谱数据为

$$[X]=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1L}\\ b_{21} & b_{22} & \cdots & b_{2L}\\ \vdots & \vdots & & \vdots\\ b_{M1} & b_{M2} & \cdots & b_{ML}\end{bmatrix}, \tag{7}$$

其中，$b_{ij}=0$ 或 1，$i=1,2,\cdots,M$；$j=1,2,\cdots,L$. 这个待识别命令 $[X]$ 跟第 k 个参考音 $[SBP]^{(k)}$ 之间的差距用下式计算：

$$D(k)=\sum_{i=1}^{M}\sum_{j=1}^{L}[\![a_{ij}^{(k)}+b_{ij}]\!]\qquad k=1,2,\cdots,B, \tag{8}$$

其中$[\![\cdot+\cdot]\!]$指进行异或. 从这 B 个 $D(k)$ 中选出最小者，例如 $D(K)=\min\{D(k),\ k=1,2,\cdots,B\}$，则识别结果便是字表中的第 K 个项目.

五、实验结果

实验是在计算机房即席呼音进行的，机房的环境噪声为 69 分贝（*A*）. 传声器用的是 CH11 型国产电容式传声器，发音时传声器位于嘴的侧前方约 5 厘米. 下列参数中，TDS 代表一个频谱所用的滤波通道数目；SPS 代表每个参考音用的频谱数目；STDH 代表频谱的首通道代号（12—160 赫兹，13—200 赫兹，14—250 赫兹等）. 滤波通道是连续地选用的.

为了适应今后连呼语言的识别研究，每个滤波器输出处的检波时间常数用了较小的数值，即 10 毫秒.

实验 1

发音人 Y（男）； 材料 0, 1, …, 9. 呼 100 遍（电报呼号）；参数 SPS＝10, TDS＝16, STDH ＝ 12；正确率 100%；错误 0/1000.

实验 2

本实验是**实验 1** 的继续，时隔 40 天，用 40 天前的参考音，继续呼十个数字，呼 100 遍所得的结果是：正确率 99.4%；错误 6/1000. ＜9＞错成＜0＞，6 次. 错误全部出现在最后二十遍中，发音人感到嗓子有点受不了. 与**实验 1** 合计，正确率是 99.7%.

实验 3

发音人 Y（男），用家乡音（吴语）；材料 20 句话（每句七字）；参数 SPS ＝ 20, TDS ＝ 16, STDH ＝ 14；正确率 99.7%；错误 （1）3/1000，（2）3/1000.

实验 4

发音人 Y（男），用家乡音（吴语）；材料 100 个四字成语；参数 SPS＝20, TDS＝16, STDH ＝ 14； 正确率 99.3%；错误 （1）第一天 0/599，（2）第二天接着发音 13/1301.

实验 5

发音人 B（女），北京人用普通话；材料 100 个四字成语；参数 SPS＝20, TDS＝16, STDH ＝ 14；正确率 99.7%；错误 （1） 2/700，（2） 4/1000.

实验 6

发音人 Y（男），用家乡音（吴语）；材料 150 个四字成语；参数 SPS＝20, TDS＝16, STDH ＝ 13；正确率 99.3%；错误 3/450.

实验 7

发音人 Y（男），用家乡音（吴语）；材料 200 个四字短语；参数 SPS＝20, TDS＝16, STDH ＝ 13；正确率 98.8%；错误 （1）第一天发音 7/800，（2）第二天接着发音 13/800，（3）第二天下午接着发音 3/400. 总计错误 23/2000.

实验 8

发音人 Y（男），用家乡音（吴语）；材料 400 个四字短语；参数 SPS＝16, TDS＝16, STDH ＝ 13；正确率 97.7%；错误 18/800.

实验 9

发音人　Y(男)；材料　0，1，…，9(电报呼号)；参数　SPS=10，TDS=16，STDH=14.

试验发音持续时间变动范围．下面的数据是发音持续时间，单位是毫秒．这是从一组试验中挑选出来的能正确识别的长音与短音的例子．

口　呼　数　字

	0	1	2	3	4	5	6	7	8	9
长　音	360	450	360	490	390	730	480	800	740	260
短　音	120	170	170	160	110	160	190	220	150	190

另外，在来宾参观时，曾做过一些非正式实验．例如，用英语呼 zero, one,…, nine; zero, one,…, nineteen; 二十条 basic 语句名字 READ, DATA, DIMENSION, MATRIX, LET, FOR, NEXT, IF, THEN, GO TO, 等；都得到满意的结果．还曾用 40 条指挥水下机动模型的命令进行试验，其中既有单音节的数字，也有两字、三字、乃至五字一句的命令，少量的几遍呼音，正确识别率也在 99% 左右．

六、讨　论

1. 从频谱的声刺激概念出发，我们用线性办法解决非线性时域规正参数的办法，再次证明是可靠的．经过对语言识别相当一段时期的工作，体会到频谱的声刺激概念很可能会对不均匀信息率声码器中参数发送时刻的选择作出贡献．这种规正参数办法适用于识别发音人．

2. 二值频谱法不仅适用于少量的口呼数字或几十个字的小字表，只要对量化时所用的阈值选取好，对于较大规模的字表也完全适用．我们对频谱分量进行二级量化时选用频谱的平均值作为量化阈值，有其优越性：即简单(一个频谱一个阈值)而且是浮动的(随频谱而变)．可以认为，我们对频谱所作的二值化处理，在克服音量变化所产生的影响方面作出了贡献．此法还具备一定的抗噪声能力，因为当某个通道里噪声低于量化阈值时，就不会产生干扰．

3. 我们的系统之所以有较高的正确识别率，就在于：(1)参数的时域规正方法抓住了语音变动的特点；(2)二值化处理对于幅值规正很成功．

4. 方案还有欠缺的地方，没有频域上的规正，尤其没有利用音调信息，而音调信息对我们汉语又是非常重要的．

5. 我们的方案，训练机器很简单，只要用户呼一遍字表(命令表)．换人或换字表都是这样．所以使用方便．而且，如果用户的家乡话能说得很稳定的话，则完全可以用家乡话来发音，这样就能得到广泛的应用．

我们称 RTSRS(01) 系统是个通用系统，一方面是指，它能‘适应’发音人的家乡话，另一方面是指，它又能‘适应’不同的字表，甚至外国的语言也行．

6. 我们用的计算机是小型通用计算机 Varian 620/L，存取周期是 1.8 微秒，这在小

型机中是慢速的，如果用 Varian 620/L-100(存取周期为 0.95 微秒)这样的 1 微秒级的标准小型计算机，那么我们的 RTSRS(01) 系统完全能实时识别字表大小为 400 这样规模的命令组. 如果用硬件来实现我们的系统，实时能力将提高二十倍以上.

表 1 为 RTSRS(01) 系统与国外同类先进系统(单呼字的识别系统)[4-7]主要参数的比较.

表 1 RTSRS(01) 与国外同类系统比较

作者	字表	正确识别率(%)	每个样板音(以600毫秒计)所需存贮量	反应时间	计算机	环境噪声级(分贝)
Martin	10个英语数字	99.79 (10人平均)	512 比特	接近实时	SIGMA-3	90(A)
(1975)	132 条机场命令	99.32 (10人平均)		接近实时		
	34 个英语数字 1—34	98.5 (12人平均)		接近实时		
White	36个英语字母及数字	98.0 (单人)	7200 比特	30倍于实时	NOVA	65(A)
(1975)	91 个北美州名	99.6 (单人)		—	PDP11	
Itakura	36个英语字母及数字	88.6 (单人)	4480 比特	—	DDP516	68(A)
(1975)	200 个日本地名	95.65 (单人)		22倍于实时		
物理研究所	10 个数字	99.7 (单人)		实时		
(1975—1976)	20 句话(每句 7 字)	99.7 (单人)		实时		
	4 字短语:					
	100 条	99.5 (2人平均)	<160 比特	实时	Varian620	69(A)
	150 条	99.3 (单人)		实时		
	200 条	98.85 (单人)		实时		
	400 条	97.7 (单人)		接近实时		

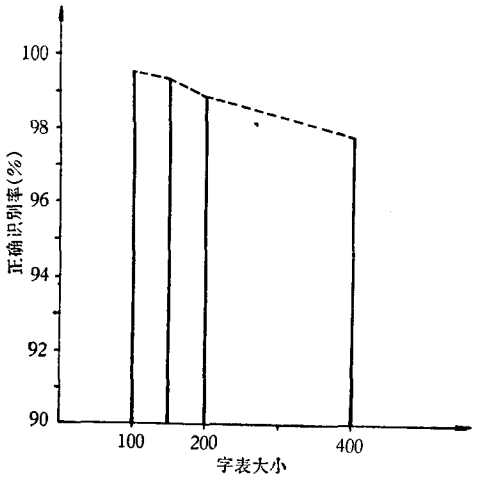

图 3 四字短语正确识别率随字表大小的变化情况

7. 图 3 中画的是对四字短语的正确识别率随字表大小变化的曲线. 可以看出，我们的系统识别能力不随字表的加大而快速下降. 由于字表加大是在原字表基础上扩大的，可以得出结论，实验结果比较稳定.

8. 由于 RTSRS(01) 系统具备实时、通用、需要的存贮量小、识别率高而且稳定等优点，已经可以付诸实用.

参　考　文　献

[1] 俞铁城，物理学报，**26** (1977)，389.
[2] L. Gilli and A. R. Meo, *Acustica*, **19**(1) (1967), 38.
[3] G. Ruske, Proc. 8-th ICA, vol. 1, 243.
[4] F. Itakura, *IEEE Trans. on ASSP*, **ASSP-23**(1) (1975), 67.
[5] D. R. Reddy, *Proc. of the IEEE*, **64**(4) (1976), 501.
[6] T. B. Martin, "Speech Recognition: Invited Papers of the IEEE Symp." New York: Academic Press (1975), 55.
[7] G. M. White and R. B. Neely, *IEEE Trans. on ASSP*, **ASSP-24**(2) (1976), 183.

A UNIVERSAL REAL-TIME SPEECH RECOGNITION SYSTEM ——RTSRS(01)

YU TIE-CHENG
(*Institute of Physics, Academia Sinica*)

ABSTRACT

In this paper a universal real-time speech recognition system-RTSRS(01) is described. On the basis of the previous work[1], the parameters of a spoken command are normalized in the time domain. Using the binary spectrum[2,3] as the final recognition parameters, which can largely reduce the amount of memory necessary for each reference command, and adapting a new method for calculating the separation between two spoken commands to be compared, it is possible to make the system RTSRS(01) capable of identifying single entities in a vocabulary of 200 items in real-time. The results of recognition for a specific speaker are as follows: 10 spoken Chinese digits — 99.7%; 20 sentences (7 syllables for each)—99.7%; 100 phrases (4 syllables for each) — 99.5%; 150 phrases (4 syllables for each) — 99.3%; 200 phrases (4 syllables for each) — 98.8%; 400 phrases (4 syllables for each) — 97.7%. It has been shown by informal experiments that the system RTSRS(01) can be used to identify the vocabularies which include items with different numbes of syllables; furthermore, for the first 20 English digits and the names of BASIC statements, the correct recognition rate is also high.