采用相关矩阵识别口呼数字的实验*

中国科学院物理研究所语言识别组

随着科学技术的发展，生产自动化程度愈来愈高，为此要求人们能以最方便的方式进行人机通信，实现语言直接指挥机器运转、排版、计算和记录等操作过程. 由于人们对于语言的产生和接收的生理过程及语言特性等了解得还很不够，现在还不能做成一种机器来识别人的全部语言，只能做到识别一些特定的专用词汇. 目前对少量词汇的识别，国际上搞的方案很多，多数方案需要快速电子计算机和相关处理等外围设备，而做成硬件付诸应用的却很少. 我们自 1971 年以来研究汉语口呼数字的识别，做过共振峰提取的实验，在计算机上模拟图样匹配方法等，去年又作了相关矩阵方法的实验. 这些工作的目的是一方面摸索汉语的声学特性，另一方面试图获得有限的应用. 下面简单介绍相关矩阵方法的原理和实验技术.

一、原理和设计

数字音的动态频谱可以表示成一个三维矩阵 $[A]_d$，$a_{t,f,d}$ 表示矩阵 A 的元素，下标 t，f 和 d 分别表示时间维的样谱数，频率维的通道数和十个数字 1 2 … 9 0；由于人的发音强弱差别较大，需要在十个音之间进行强度规正，规正化的系数按下式计算：

$$b_{t,d}=\left(\frac{1}{10}\sum_{d=1}^{10}\sum_{f=1}^{n}a_{t,f,d}^2\Big/\sum_{f=1}^{n}a_{t,f,d}^2\right)^{\frac{1}{2}}.$$

规正化后的矩阵表示为 $[X]_t$，它的元素为

$$x_{d,f,t}=a_{t,f,d}\cdot b_{t,d}.$$

定义 $[X]_t$ 和 $[A]_m$ 之间的相关函数为

$$c_{d,m,t}=\sum_{f=1}^{n}x_{d,f,t}\cdot a_{t,f,m}.$$

$[A]_m$ 是任意一个数字阵，表示数字的下标用 m 以区别 $[X]_t$ 中表示数字的下标 d. $[C]_t$ 是以 $c_{d,m,t}$ 为元素的相关矩阵，由 $[X]_t$ 和 $[A]_m$ 计算 $[C]_t$，最后累加比较就得到以 d 识别 m 的结果.

为了预计这种方法的效果，可以先在计算机上模拟计算，我们选择 1/3 倍频程滤波器的中心频率从 200 Hz—4 kHz，共十四个通道，样谱间隔 20 ms，只用前十个谱，选取 50 次（每个数字重复 5 次）作为平均的 $[X]_t$，对 100 个 $[A]_m$（每个数字重复 10 次）进行计算，累加 40 ms 的结果与选择样谱的关系如表 1 .

表中数字是每个样谱对每个音呼十次认对的次数. 可见不同时刻的样谱对每个音的

* 1976 年 3 月 26 日收到.

表 1　各样谱在 40 ms 内的识别率

选择样谱		1	2	3	4	5	6	7	8	9	10
每音呼10次	1	3	5	9	9	10	8	9	9	7	9
	2	6	8	10	9	9	6	7	10	8	8
	3	8	9	10	10	10	10	10	10	10	10
	4	2	7	6	7	9	9	9	6	5	2
	5	10	10	10	10	9	8	8	8	1	1
	6	5	6	3	9	10	9	8	9	8	9
	7	3	8	8	10	10	10	9	10	9	10
	8	10	10	10	10	10	10	10	10	10	9
	9	0	8	10	10	9	10	9	9	8	10
	0	10	9	7	7	6	8	7	8	9	9

辨认贡献大小不等，识别率较高的样谱对应于一个音区别于其他音的主要特征，也是该音能量较强的区域(共振峰)．我们的实验是应用 70 ms 附近的样谱制成电阻矩阵(图 1)，各电阻值按下式计算：

$$R_{d,f}=\frac{L-\sum_{f=1}^{14}x_{d,f}}{x_{d,f}}\cdot R_{\Sigma d},$$

其中 $R_{\Sigma d}$ 视下级输入电路要求而定，L 是一个标称值，它的大小决定矩阵网路的传递系数，由下式估算：

$$L\doteq 10x_{d,f\max}+\sum_{f=1}^{14}x_{d,f},$$

其中 $x_{d,f\max}$ 是 $[X]_t$ 中最大的一个元素．

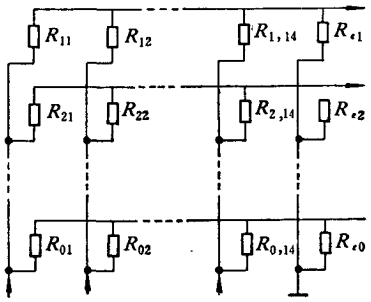

图 1　电阻矩阵

二、实验

图 2 是我们实验的方框图．先由传声器将声转换成电信号，经放大、滤波和低通网路检出其包络送矩阵；通向矩阵的道路受门电路控制，它的作用是从发声开始关闭 70 ms 后再打开进入矩阵的通道，当有了判别后即刻又关闭了通路，直到音末再开启以待接收下一个声音．由矩阵出来的信号积累 30 ms 后送检出器提取极大值输出，输出可打印或显示．

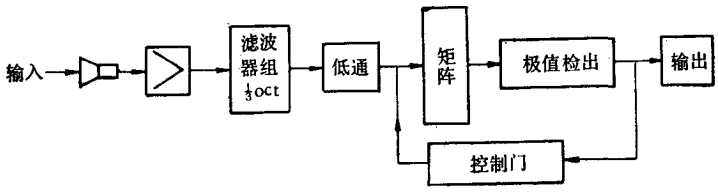

图 2　方框图

三、结果和讨论

按设计的矩阵参数识别正确率有 90% 左右，后经调正识别率可提高到 95% 以上，调正主要是提高高频峰值作用，并以 8 kHz 通道代以 3.15 kHz 和 4 kHz 二通道．因为矩阵

按专人设计，换人识别率就降低. 由于用了控制门电路二次关闭矩阵通道，前一次是调定的，后一次随发音长短而变，这就消除了发音持续时间长短的影响，并能做到实时输出，虽然仍需单呼，但发音速度可提高到每秒 3 个音节.

这种方法如设置二个样谱矩阵，经适当延时相加可有更好的结果，可以用调换矩阵参数解决换人问题. 它识别的声音可不限数字，可以是语音[1]，还可以是工业生产上的声音，从而用作工业生产中的声控制.

参 考 文 献

[1] H. Dudley, S. Balashek, *J. A. S. A.*, **30** (1958), 733—739.

THE USE OF CORRELATION MATRICES IN SPOKEN DIGITS RECOGNITION

SPEECH RECOGNITION RESEARCH GROUP,
INSTITUTE OF PHYSICS, ACADEMIA SINICA