语音识别的二值化时频图型模糊匹配法

戎 月 莉
（同济大学 上海 200092）

1993年5月26日收到

将模糊逻辑应用于语音识别系统，具有减少数据量和计算量，提高语音识别率的优点．本文阐述了二值化时频图型模糊匹配法（BTSP）的原理，并对它目前的一些应用产品作了简单介绍．

ABSTRACT

Using fuzzy-logic to speech recognition system reduces a large number of data, simplifies calculation and increases the successful rate of correct recognition. This paper deals with the principles of Binary Time Spectrum Pattern Matching (BTSP). Some of its recent applications are presented.

一、引 言

语音识别方法，有隐马尔可夫模型技术（HMM）等．一般首先建立标准的语音特征模板，通过计算输入语音与样本语音之间的距离测度，进行反复多次校核．我们知道，语音是声压随频率及时间的二维变化过程，因为以下两种因素：1.发声韵母长度的不同和发声速度不同带来的时间变动；2.不同人的发声器官性质上的差异带来的发同一语音时频率上的差异，即频谱变动，这一动态过程呈现复杂的形态，增加了语音识别的难度．尽管现代计算机具有高速的运算能力和记忆能力，但是人工语音识别系统是远远不及人耳那样灵敏的，这是因为人的大脑具有一种模糊识别和判决能力，是目前的计算机所不具备的．如何能将智能识别能力赋予计算机，使它也具有一定的模糊模式识别能力，以提高语音识别的正确性，缩短识别所需的时间，是语音识别深入进行所面临的一大问题．本文简扼阐述一种二值化时频图型模糊匹配法（BTSP法），它将模糊逻辑引入了语音识别系统，已在实际应用中取得很好的效果．

二、模糊逻辑的引入

1965年美国自动控制专家扎德（L. A. Zadeh）提出了他有关模糊逻辑的第一篇论文《Fuzzy Set》（模糊集合）时，就首次提出"隶属度函数"的概念．它给出了模糊概念，即那些没有绝对分明的界限的概念的定量表示．一个模糊集合 A 是通过隶属度函数 $m_A(x)$ 来定义的：

$$m_A=\begin{cases}1 & \text{表示 } x \text{ 完全属于集合 } A\\ 0 & \text{表示 } x \text{ 完全不属于集合 } A\\ \text{其它值} & \text{表示 } x \text{ 属于集合 } A \text{ 的程度}\end{cases}$$

这里，$m_A(x)\in[0,1]$．这样，就将普通集合进行了扩展．普通集合中只有 $m_A(x)=1$ 与 0 两种可能，故和计算机的二值逻辑相对应．模糊集合论的出现将二值逻辑扩展到取值为[0,1]闭区间任意值的连续值逻辑[1]．因此模糊逻辑的重要特点是它吸取了人脑识别和判决的模糊性，使它运用于计算机，让计算机也能模拟人的思维，以完成复杂的任务，这正是模糊逻辑出现的背景．

仿照普通集合中的并、交、补等基本运算，在模糊集合中也有相应的逻辑运算．若以 A、

B、C 表示三个模糊集合，则

当 C 为 A 与 B 的并集，即 $C = A \cup B$ 时，

$$m_C(x) = \max[m_A(x), m_B(x)]$$

当 C 为 A 与 B 的交集，即 $C = A \cap B$ 时，

$$m_C(x) = \min[m_A(x), m_B(x)]$$

当 $\bar{A}$ 为 A 的补集时，

$$m_{\bar{A}}(x) = 1 - m_A(x)$$[2]

可见模糊集合的运算实质上是隶属度函数的运算过程。

三、模糊模式匹配法

1. 模糊模式匹配原理

由于发声单词在时域和频域上存在变动，我们认为输入的单词语音图型具有模糊性。若语音识别系统能识别的单词共有 n 个，定义这些单词的集合为 I，单词图型的集合为 X，而其隶属度函数的集合为 M。可得

$$I = \{i_1, i_2, i_3, \cdots, i_n\}$$

$$X = \{x_1, x_2, x_3, \cdots, x_n\}$$

$$M = \{m_1, m_2, m_3, \cdots, m_n\}$$

这里的 X 是模糊集合[3]。

我们先从一维的情况，即认为输入的未知语音图 y 只是频率的函数进行讨论。图 1 中的横轴表示频率。图 1(a) 表示的是单词 i_j 的隶属度函数，它在共振频率所在的位置，假定是单峰的。图 1(b) 表示的是输入的未知语音的图型，也以共振频率表示。那么未知语音属于单词 i_j 的程度（即隶属度）d 可由隶属度函数 m_j 与 y 的交求出，

$$d = y \wedge m_j$$

因为语音的共振频率常常不限于一个，因此还

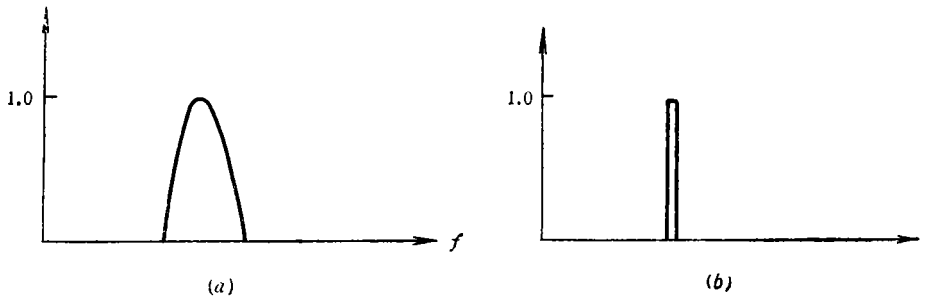


图 1　隶属度函数和输入图型

须使用补集。单词 i_j 的补集的隶属度函数以 $(1 - m_j)$ 表示，那么表示输入的未知语音 y 不属于单词 i_j 的程度 $\bar{d}$ 为

$$\bar{d} = y \wedge (1 - m_j)$$

定义以上两式之比为单词 i_j 与未知语音 y 的相似度 S_{jy}

$$S_{jy} = \frac{d}{\bar{d}} = \frac{y \wedge m_j}{y \wedge (1 - m_j)}$$

实际上输入的未知单词图型和隶属度函数都认为是二维的，因此有

$$d_{jy} = \sum_f \sum_t m_j(f,t) \wedge y(f,t)$$

$$\bar{d}_{jy} = \sum_f \sum_t [1 - m_j(f,t)] \wedge y(f,t)$$

$$S_{jy} = d_{jy} / \bar{d}_{jy}$$

其中 f 表示频率，t 是时间变量。

根据模糊逻辑中关于模式识别的择近原则，应取相似度 S_{jy} 最大的单词 i_j 作为语音识别结果。若 J 表示与输入未知语音 y 相似度最大的单词 i_j 的序号，且记为

$$J = \max\{S_{jy}\}$$

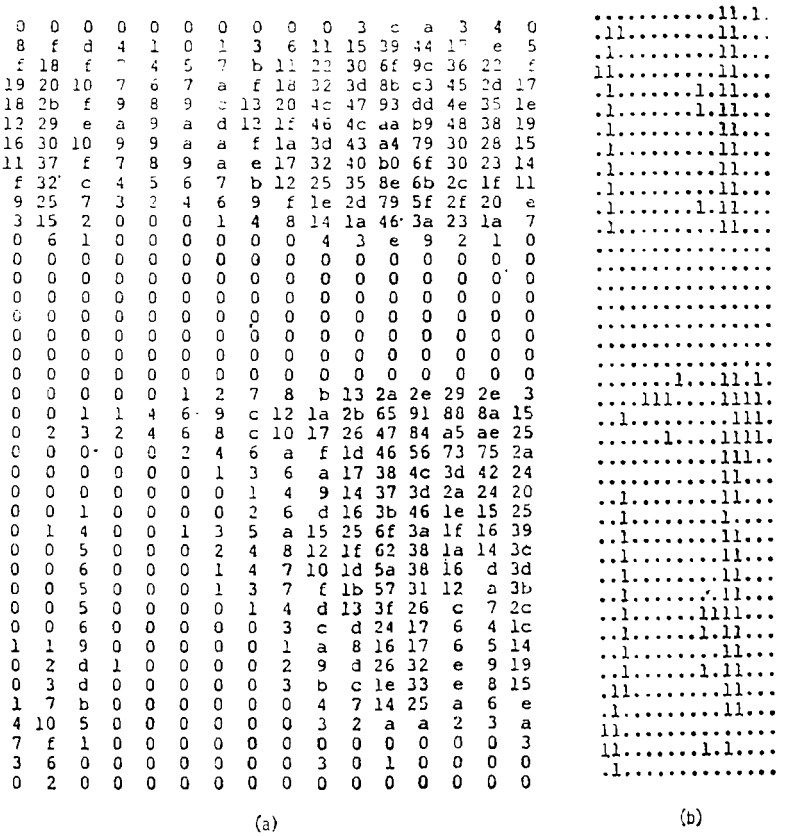
J 就可以决定识别的结果。

这种用单词 $i_j (i_j \in I)$ 的隶属度函数 m_j 与未知语音进行模糊逻辑运算，求出各单词与未知语音 y 的相似度 $S_{1y}, S_{2y}, \cdots, S_{ny}$，取其中最大值相应的单词作为语音识别结果的方法就是模糊匹配法。

2. 隶属度函数的确定

上述的单词的隶属度函数，作为语音识别中单词的标准样本，又是怎样确定的呢？

以往在语音识别中，作为校核标准的单词，可以 TSP 图 (Time Spectrum Pattern) 表示，就是把声波在短时间间隔进行频率分析而得到的。例如对某话声者发出“1”(日语)的声音，从最低频率 250Hz 到最高频率 6500Hz 分成 16 个频带，每 10ms 一帧作短时谱分析。图 2(*a*)所示的就是用 16 进制表示的单词“1”的 TSP 图。



(a)　　(b)

图 2　TSP(*a*)和 BTSP(*b*) 图(单词名“1”)

如果直接以 TSP 图作为样本，那将是极复杂的。由于同一语音本身因时间变动和频率变动产生的多样性，又因为 TSP 图上每一元素的值以 12 位 (Bit) 表示，数据量和计算量相当大，而显示特征的共振频率却并不突出，因此发展了一种新的 BTSP 法 (Binary Time Spectrum Pattern)。它是将 TSP 图进行二值化，我们将图 2(*a*) 每一行的峰值取作 1，(在图上以“1”表示)，其余值全部取作零(在图上以“·”表示)，就得到了图 2(*b*)，即单词“1”的 BTSP 图*。可见，BTSP 图既保留了并突出了发声单词的特征，即共振频率，又大大缩减了数据量，因为 BTSP 图上每一元素仅占 1 位 (Bit)。我们把这种 BTSP 图称为单词图型。

但是，不同的发声者所发的同一单词的 BTSP 图是不相同的。因此还须将许多不同人发同一单词语音的 BTSP 图进行处理，例如线性地伸缩长度，进行反复重合操作，才能得到最终的作为校核标准的单词特征图型，它就是进行模糊模式匹配所需要的隶属度函数。

---

* 只要满足下列条件之一就可取作峰值

1) 一行中的某一频带，其上值比左右频带上的值都高。如第一行末尾…340，其中 4 即是峰值。

2) 一行中的某一区域(可包括 *n* 个频带)，其上的值比左、右频带的值高出许多，而它们相互相差不大。如第一行中…3*c*a3…，其中的 *c*、*a* 处取为峰值。(16 进制中 *a* 等于 10，*c* 等于 12)

3. 应用模糊模式匹配的语音识别方法

利用模糊模式匹配方法进行语音识别的流程见图 3.

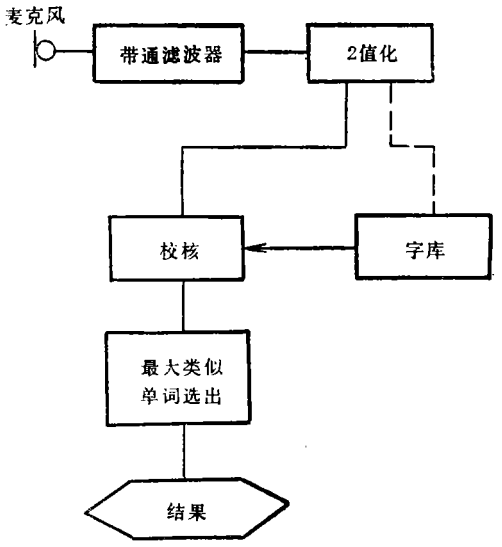


图 3　模糊模式匹配法识别语音的流程图

输入的语音信号经过带通滤波器后，进行二值化，即作成 BTSP 图. 预先将经过处理作为标准的单词图型(即隶属度函数)存入字库. 输入的未知语音的 BTSP 图，按前述的模糊模式匹配方法与字库中标准的样本进行校核，从中选出相似度最大的单词作为识别结果[4].

这一语音识别系统无论是对特定话声者，还是不特定话声者都是适用的. 若使用于特定话声者的场合，只要按特定话声者发的几次声音来"训练"隶属度函数即可，它能得到更高的语音识别率.

四、应用与设想

应用上述二值化时频图型模糊匹配方法进行语音识别的装置，在国外现在已达到商品化阶段.

在 1987 年国际模糊系统会议上 (IFSA' 87)，日本已出售过语音识别装置 RV100 (适用于特定话声者)和 RV100I (适用于不特定话声者)，它们识别的单词数为 120 个. 现代的声音认识电话，使人们可以不必查询和记忆电话号码，只要拿起电话耳机，呼叫对方的名字就可以接通电话. 它能记录与 120 个人名对应的电话号码，这在 1987 年东京的商业展览会上已展出过. (目前已进入我国市场) PV-VOICE RECOGNIZER 是一种语音识别的插件板，板上有特制的大规模集成电路，用 BTSP 法识别语音. 它有二种类型，分别适用于 IBMPC/AT 机和 NEC 的 PC9800 系列. 当将此板插入 PC 机的扩充槽后，就能进行数据的传送，而计算工作则由 PC 机的 CPU 完成. 它的特点是把语音识别的控制软件固化在 PC 机中，详见参考文献[3].

新近出现的由 Ricon 制造的"SWR-U3-02"语音识别单元[5]，则利用二个声频接收通道，将输入信号频谱相减来达到有效抑制背景噪声的目的，见图 4. 这一系统能识别的指令字有 120 个，平均每个字长为 1 秒. 这在工业应用范围内已经足够了. 它已在汽车上进行过试验，当汽车在高速公路上以 140km/小时的速度飞驰时，噪声相当大，然而也可得到极好的识别效果. 利用模糊逻辑方法能精确地识别出所述的命令，可见这一语音识别法具有较高的鲁棒性.

二值化时谱图型模糊匹配法是在语音识别中引入模糊逻辑方法而产生的. 事实证明这一方法对语音识别很有效. 我们考虑，在声学的其它问题上，如声源的识别、噪声等方面是否也可以应用模糊逻辑方法. 结合人耳的生理构造，模仿人耳，设计二个声频接收器，然后再用

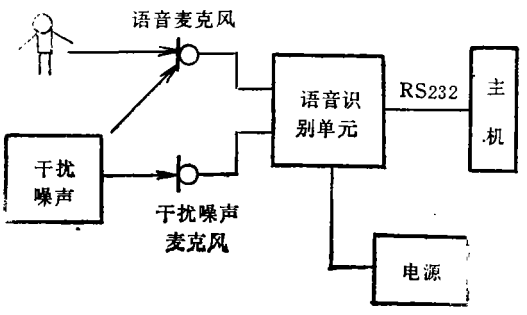


图 4　语音识别系统方块图

模糊逻辑方法进行信号处理．目前世界上流行的神经网络热，正与模糊逻辑结合起来，可用神经网络“训练”隶属度函数．这些是否也可以应用到声学领域，使得人工语音识别系统达到较灵敏的水平，还有待于人们的努力．

参 考 文 献

[1] 王学慧，田成方，微机模糊控制理论及其应用，电子工业出版社，1987.

[2] 汪培庄，模糊集合论及其应用，上海科学技术出版社，1983.

[3] 藤本潤一郎，ファジイの音声認識，コニピエートロール，35(1991)，78—84.

[4] 藤本潤一郎，ファジイパターンマッチングにもぬ音声認識，電子技術，1(1991)，43—47.

[5] Ha, Rechner hört mit Fuzzy-Logik, Elektronik, 17(1992), 16.

数 字 音 响 系 统（二）

沈 蠔

（中国科学院声学研究所，北京 100080）

1994 年 1 月 13 日收到

三、数字式唱片放声系统

1. 数字音频唱片

由于 PCM 信号的频带很宽，要直接以 PCM 信号记录在唱片上需要有新的高密度记录方式．若 PCM 信号采样频率为 45kHz，13bit，双通路，则要求带宽约为 1.2MHz，是一般密纹唱片的 60 倍．若要记录四通路立体声信号，则要求 2.4MHz 带宽．因此采用可记录 10MHz 信号的激光视频唱片为记录媒质．七十年代发展的数字音频唱片（DAD）有机械式，静电式和激光式三类．机械式 DAD 最简单，刻纹和重放方式与普通唱片相似．放声时采用压电换能器，拾取波长为 0.5μm 的信号．解调后的数字音频信号经误码校正处理，由 D/A 转换器还原为模拟音频信号．静电式 DAD 采用激光烧蚀刻录，激光束的强弱由光调制器控制，按信号的强弱使光敏抗蚀剂感光，然后显影留下与信息相应的凹坑．重放时采用无导向电容针读取方式，多边形宝石唱针与唱片表面接触，兰宝石唱针一面带有电极；由于唱片是导电性的，因此唱片与电极之间构成一个电容，其电容量随信息坑的有无而变化．唱片与唱针电极之间的电容与线圈构成谐振器并与 UHF 振荡器耦合．电容变化时谐振频率就变化，输出电平也随之变化，从而读出唱片上信息．它的特点是和现有静电式数字电视唱片兼容．唱片用 PVC 导电材料压制，表面刻有细密的主信息凹坑和跟踪信号凹坑．唱片上记录三条通道的音频信号和一条通道的静止画面信息．既可供给普通双通路信号，也可供给三通路立体声信号，并且还可在欣赏音乐节目的同时看到风景画面或演奏者的图片．激光式 DAD 刻纹时也用激光来烧蚀，重放时采用非接触方式，由激光束替代唱针，光电二极管作换能元件．它采用了 PCM 信号数字处理技术和激光超高密度记录技术，其性能很好，后来发展为 CD 唱片并且商品化，下一节将详细讨论．

数字音频唱片的录音和放音系统如图 3 所示．节目经低通滤波器、采样电路、模数转换和记忆电路后形成脉码调制波形，然后馈送给宽频带刻片设备录制原版唱片．这种 DAD 唱片必需用相应的宽频带唱盘来拾取信息，并经过信号分离、误差检出、记忆、数模转换、低通滤波和放大等过程，才把数字音频编码信号还原为