第四讲　语音信号处理的现状和展望*

李　昌　立

（中国科学院声学研究所　北京　100080）

摘　要　　文章简要介绍了“语音信号处理”这一分支学科形成和发展的历史过程.指出了它在现代信息科学技术中的地位和作用.介绍了语音信号处理在应用领域的一些重要课题,如语音的低速率编码,语音的规则合成和文-语转换系统,语音识别和人-机语音对话等,这些仍然是当前研究的热点.文章最后展望了语音信号处理的发展前景,指出在这个领域还有很多难题等待人们去研究探索.

关键词　　语音信号处理,语音低速编码,语音识别

Current status and prospects of speech signal processing

Li Chang-Li

(*Institute of Acoustics*, *Chinese Academy of Sciences*, *Beijing* 100080, *China*)

Abstract　The history of speech signal processing and its status in modern informatics and information technology is reviewed. In practical applications, key techniques such as low bit rate speech encoding, speech synthesis by rule, text to speech conversion, speech recognition, speech dialogue between man and machine are still hot topics for current research. Though much has been achieved in past years, there are many problems to be solved. Future developmentsof speech signal processing are identified.

Key words　speech signal processing, low rate speech coding, speech recognition

1　简要的历史回顾

声学是物理学的一个分支学科,而语言声学又是声学的一个分支学科.它主要的研究方向是人的发声器官机理,发声器官的类比线路和数学模型,听觉器官的特性(如听阈、掩蔽、临界带宽、听力损失等),听觉器官的数学模型,语音信号的物理特性(如频谱特性、声调特性、相关特性、概率分布等),语音的清晰度和可懂度等.当今通信和广播的发展非常迅速,而语言通信和语言广播仍然是最重要的部分,语言声学则是这些技术科学的基础.

语言声学的发展和电子学、计算机科学有着非常密切的关系.在它发展的过程中,有过几次飞跃.第一次飞跃是1907年电子管的发明和1920年无线电广播的出现.因为有了电子管放大器,很微弱的声音也可以放大,而且可以定量测量.从而使电声学和语言声学的一些研究成果,扩展到通信和广播部门.第二次飞跃应该是在20世纪70年代初,由于电子计算机和数字信号处理的发展,人们发现:声音信号特别是语音信号,可以通过模数转换器(A/D)采样和量化,它们转换为数字信号后,能够送进计算机.这样就可以用数字计算方法,对语音信号进行处理和加工.例如频谱分析可以用傅里叶变换或快速傅里叶变换(FFT)实现,数字滤波器可以用差分方程实现.在这个基础上,逐渐形成了一门新学科——语音信号处理.它的发展很快,在通信、自动控制等领域,解决了很多用传统方法难以解决的问题.在信息科学中占有很重要的地位.

2　语音信号处理在信息科学中的地位和作用

众所周知,语音在人类社会中起了非常重要的作用. 在现代信息社会中,小至人们的日常生活,大到国家大事、世界新闻、社会舆论和各种重要会议,都离不开语言和文字. 近年来,普通电话、移动电话和互联网已经普及到家庭. 在这些先进的工具中,语音信号处理中的语音编码和语音合成就有很大贡献. 再进一步,可以预料到的口呼打字机(又称听写机,它能把语音转换为文字)、语音翻译机(例如输入为汉语,输出为英语,或者相反),已经不是梦想而是提到日程上的研究工作了. 人们早就希望用语音指挥机器,机器的执行情况也能用语音回答. 这在某些领域已经部分地实现了. 目前计算机芯片的集成度和运算能力,每 18 个月就提高一倍,而成本又不断降低,因此,它已经广泛地应用于在社会生产和生活的各个方面. 然而计算机接收信息的外围设备和主机相比,要逊色得多. 能说能听的计算机还不能普遍使用. 也就是说:语音识别、语音理解和语音合成等课题,还有很多理论问题和技术问题没有解决,需要继续深入研究.

科学家们深入研究后认为,要解决人 – 机语音对话这样的难题,做出真正实用的语音机器,必须开展跨学科的研究,如声学、语言学、语音学、生理学、数字信号处理、人工智能和计算机科学等. 要真正赋予微电脑以语言功能,必须彻底了解语言是如何产生、感知,以及人类的语言通信是如何进行的?图 1 给出了从语言产生到语音感知全过程中的几个重要环节. 从图 1 可以看到,要使这个问题得到满意的解决,需要深入研究人类发声器官和听觉器官机理,建立能反映客观真实情况的物理模型和数学模型.

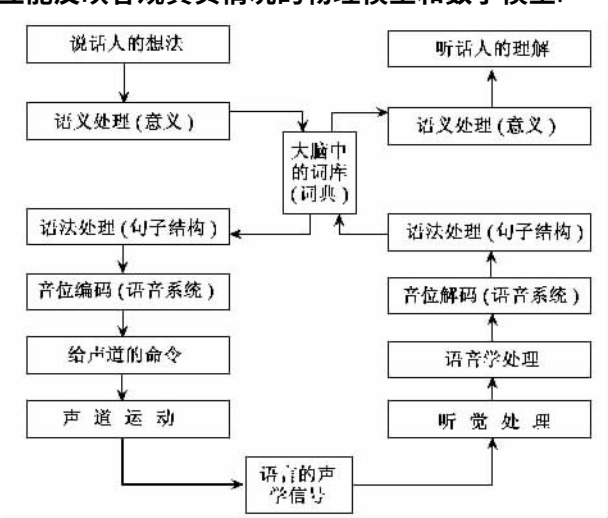

图 1 人类语音通信的过程[1,2]

3 语音信号所包含的信息量[1,2]

语音信号中到底包含了多少信息量,需要多少比特才能够无失真地表示它们,这对于语音编码、语音合成和语音识别的研究都是很有用的. 但是这也是一个很复杂的问题,它涉及到对于信号失真的评价. 下面列举了三种评价,其中两种是由弗累雷格(Flanagan)给出的,另一种是由约翰斯登(Johnston)提出的. 它们是建立在下面三种不同的失真评价上:(1)语音信号的信噪比;(2)接收语音信号时,信号由听觉外围处理以后,人们在主观上能够感觉到的失真;(3)人在接收语音信号时,不正确接收音素的数目和正确接收音素数目的比值.

在所有的三种情况下,所得到的比特率是首先选择能够接受的失真等级,然后,计算该失真等级所需的比特率. 在测量音素失真的情况下(第三种),可以把接受的失真级设置为零. 如果所有的音素都能正确传送,就是所期望的最好性能. 假设相邻的音素之间不出现相关,则平均信息速率很容易计算. 按照仙农(Shannon)的信息理论,每一个符号需要的平均比特数为

$$I = -\sum_i p_i \log_2(p_i), \qquad (1)$$

式中 p_i 为每一个符号 i 的概率,英语有 42 个音素(符号),汉语的音素有 48 个,其中辅音 22 个,单元音 13 个,复元音 13 个. 在正常情况下,谈话速率大约是每秒钟 10 个音素. 使用音素出现的相对概率表,能够计算出每一个符号的信息量大约是 5bit,得到的全部信息速率大约是 50bit/s. 请注意,自然的静寂也包含在这个比特速率内. 而系统仅仅传送音素序列,缺少发音人声音的个性特征(也就是声带的形状和对声道的描述). 在另一方面,相邻音素之间的相关也被忽视了. 考虑到这些音素后,把这一估计作为语音信息所需要的比特率低限,或者人们感知语音信号的最低要求,还是有一些道理的.

其次,把语音信号的信噪比作为失真评价(第一种),在不考虑编码器结构的情况下,可以得到语音信号信息速率的高限. 在具有电话带宽的信号中,估计最大信息速率时,必须要考虑合理的噪声等级. 令 P 是信号的平均功率,W 是信号的带宽,G 是附加的噪声信号功率,假设附加的噪声信号是高斯白噪声,令 C 表示最大的信息速率,由仙农的理论,对于包含了附加噪声 G 的信号,C 可由下式计算.

$$C = W\log_2\left(1 + \frac{P}{G}\right). \quad (2)$$

在上式中,如果语音信号的带宽为 3.5kHz,信噪比(SNR)为 30dB,则它所包含的最大信息速率为 35kb/s. 这是语音所需要的信息速率的上限. 在上面的公式中,对于语音信号所存在的短期相关和长期相关,都没有考虑. 而信号中所存在的结构性相关,就意味着冗余度. 它能够在传输之前除去,从而降低信息速率.

下面所讨论的估计,要包括人的感知和理解(第二种). 声音信号由人的听觉器官处理以后,它的信息率就降低了. 声音信号的某些特点,会由于人听觉系统的掩蔽效应而不能被注意到. 例如在一个特有频率上的低幅度纯音,可以被一个靠近该频率更响的纯音掩蔽. 在除去了人们在感觉上不能区分的特点以后,再来考虑信号的信息速率是恰当的. 如果把理解失真评价的阈值也设置到零(不能听到失真). 则首先计算语音信号的傅里叶变换,然后按频带进行计算,要求的量化器步长应该使量化噪声在掩蔽阈值以下. 掩蔽阈值和频带宽度都是建立在听觉系统知识的基础上,所得到的信息速率估计称为理解熵. 对于电话带宽的语音,理解熵估计大约为 10kb/s. 这是对于连续语音的,相当于执行透明的语音编码所需的平均速率. 上面讨论表明,人的感知和理解在语音处理中有很重要的作用.

4 发声器官和听觉器官的物理模型[1,3,4]

人类发声器官的工作情况如下:肺及有关的呼吸肌肉是能源,气流由气管呼出,首先经过声门(声带开口处). 当发浊音时,声带振动调制气流,产生一系列的离散脉冲. 被调制的气流再经过咽腔和口腔,必要时,软颚打开还有鼻腔加入. 发不同的声音,是由于口腔张开的程度不同,舌在口中的位置不同,从而使各空腔的容积发生变化,当气流经过时就要产生许多共振,最后从口和鼻以声波的形式辐射出来. 当发清音时,声带不振动,但在喉至唇通道的不同部位,有的是对气流先阻塞然后迅速打开——塞音(如 b,d,g);有的是形成一狭窄通道产生摩擦——擦音(如 f,s,h);有的兼有阻塞和摩擦——塞擦音(如 j,zh,z).

由发声器官的机理可以得到发声器官的机电类比线路. 经过简化,激励源可用脉冲串发生器和白噪声发生器模拟,而声道传输函数能够用不同截面相连接的无损声管模拟,并能简化为只有极点没有零点的数字滤波器. 简化模型如图 2 所示.

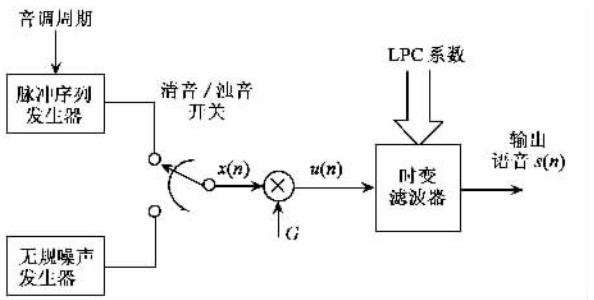

图 2 简化的语音产生模型方框图

在图 2 所表示的模型中,我们可以看到,当输入的驱动信号(或称激励信号)是由脉冲序列(对于浊音)或者无规噪声(对于清音)表示,而声门气流、声道和嘴唇的合成贡献是用数字时变滤波器表示时,则输出就是所模拟的语音信号. 这是一个重要的语音产生模型,它可以用时域公式表示,也可以用频域的公式来表示. 但在数字信号处理中,最方便的方法是用 Z 域的公式表示. $z = e^{j\omega}$ 是离散时间信号(采样序列)的一种表示方法,可以把信号映射到具有单位园的 Z 平面上,给分析和计算带来很大方便. [有兴趣的读者可以参考有关的数字信号处理教科书. 也可先不深究,只看下面时域的公式(4)和(5)]

Z 域的公式为

$$H(z) = \frac{S(z)}{X(z)} = \frac{G}{1 - \sum_{j=1}^{P}\alpha_j z^{-j}} = \frac{G}{A(z)}, \quad (3)$$

式中 $S(z)$ 是输出的语音信号 $s(t)$ 的 Z 变换,$X(z)$ 是输入激励信号 $x(t)$ 的 Z 变换,$H(z)$ 是 Z 域的系统传输函数.

把(3)式变换到信号采样的时间域,我们得到

$$s(n) = Gx(n) + \sum_{j=1}^{P}\alpha_j s(n-j). \quad (4)$$

(4)式就是著名的线性预测(LP)差分方程,在 LP 分析中,问题能够陈述如下:给出信号 $s(n)$ 的测量值,可以求出参量即预测系数 α_j,$j = 1, 2\cdots p$(具体的计算方法可参阅文献[6])如果假设所得到的参量,就是所模拟的系统函数 $H(z)$ 的参量. 则它现在的输出值 $s(n)$ 是由过去输出样品的加权之和及预测误差所确定. 预测误差(或残差)为

$$e(n) = s(n) - \sum_{j=1}^{P}\alpha_j s(n-j) = Gx(n). \quad (5)$$

在上述推导中,假设模型的信号是稳定的,也就是它只适合于长度为 20—30ms 的语音信号. 因此语音

信号的预测系数要分段计算,它们是随时间而变化的. 而且可以转换成多种形式. 例如反射系数 k_i、倒谱系数 C_i、线谱频率 LSF_i 等. 它们对于语音编码、语音合成和语音识别都有很重要价值.

听觉器官的物理模型涉及到人们对声音的感知和理解. 目前,还没有像发声器官物理模型那样成熟的研究成果. 虽然也发表了不少文章,认为听觉器官可以用一组小于临界带宽的带通滤波器再加一个非线性处理器来模拟. 然而都还没有公认的物理模型和一套实用的算法. 但是,人们从听觉的研究中,对于一些容易由实验证明的问题,已经有了明确的结论. 例如在响度和响度级的实验中,证明了人耳对于不同频率的声音具有不同的灵敏度. 人耳对幅度的感觉是非线性的. 在听觉掩蔽实验中,证明了只有在纯音两旁的窄带噪声才能掩蔽纯音,换句话说,人耳有滤波器的功能. 如果噪声级用它的声谱级表示,也就是用 1Hz 带宽内的有效声压级表示,则一个可以被听见的纯音,其声压级应略大于噪声在临界带宽内的总有效声压级. 因此,当噪声掩蔽纯音时,以纯音频率为中心的某一频带内的噪声总功率起作用,这个频带宽度就称为临界频带宽度(假设噪声频谱在该纯音附近比较平). 假设某一频率的纯音和一窄带白噪声混合,还假设噪声的带宽 $f_a - f_b = \Delta f$ Hz,可以环绕纯音频率为中心作较大变化. 随着带宽的增加,掩蔽 K(以 dB 表示)将随着带宽的对数值成正比增加,直到它的临界带宽 Δf_c 为止. 超过这个临界带宽后,对纯音的掩蔽将不再增加. 由此可见,在背景噪声中倾听一个纯音信号时,如果要用带通滤波器来消除背景噪声,它的带宽必须小于临界带宽,否则并无好处. 这些概念和理论已经在工程和技术领域得到了应用.

5 语音信号的中、低速率编码[1,5,6]

按照语音产生的简化模型(见图 2),可以构成低速率的语音编码器(又称声码器). 最早的模拟声码器和以后的数字声码器 LPC - 10、LPC - 10e 都是根据这个模型设计的. 激励源使用二元激励,在同一时间只能用一种激励方式,即白噪声或脉冲串. 声道传输函数可用一组带通滤波器模拟,在更多的情况下,是把声门脉冲形状、嘴辐射和声道等因素结合起来,用一个全极点滤波器模拟. 因为人的发声器官是机械系统,运动缓慢,传送这些慢变化的控制参量,可以用速率比较低的数码. 它和传送波形所需要的数码相比,能够压缩许多倍,不但节约了频带,而且有利于保密. 在第二次世界大战中,美国和德国都使用过这种保密电话. 随着电子技术的进步,这种声码器经过精心设计和不断优化,在 2.4kbit/s 的速率下,可以产生完全可懂的语音. 美国军方和北大西洋公约组织一直用作保密电话. 但音质和自然度很差,其原因是二元激励模型有局限性,不符合客观实际情况.

科学家们经过深入研究,提出了合成 - 分析法(AbS),比较满意地解决了这个问题. AbS 方法并不是惟一的用于语音编码,而是估计和验证领域的通用技术. 它的基本概念如下:首先,假设产生信号模型的方式如图 3 所示. 这个模型受一些参量控制,改变这些参量就能够产生不同的观测信号. 要使所表示的模型和真正的信号模型有同样的形式,能够使用一个试探程序或误差程序,采用有规则的方法改变模型参量,从而可以找到一组参量,它所产生的合成信号,能够以最小误差与真正的信号相匹配(假设模型开始就是有效的). 因此,当计算到这样的匹配时,模型的参量就可以认为是真正信号的参量.

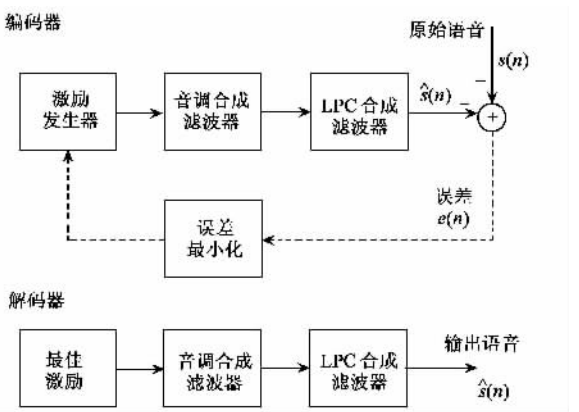

图 3 使用合成 - 分析法的语音编码方框图(采用 AbS - LPC 编码方案)

AbS - LPC 方案(使用合成—分析法的线性预测编码)的基本操作如下:

(1)将 LPC 和音调滤波器(时 - 变滤波器)的内容,初始化到预定的值(通常是置到零或低量级无规噪声).

(2)缓冲一帧语音样品,在该帧上使用 LPC 分析算法,计算出一组 LPC 系数.

(3)使用计算得到的 LPC 系数,构成一个反滤波器,计算非量化的残差信号.

(4)为了有效地分析激励信号,把分析帧再分为整数子帧. 对于每一个子帧:(i)计算音调滤波器

(长期预测器)的参量,也就是延迟 τ 和与它联系的标量因子 β. (ii)按照图 3 中的级联滤波器,则最优的辅助激励可以按照合成语音和原始语音之间的最小误差方法确定.

(5)最后的合成语音,是由最优辅助激励通过具有初始存储内容的级联滤波器产生的(初始存储内容是从以前子帧合成过程中残留下来的).

这种方案运算量很大,但话音质量好,数码率也可以做得很低(16k—4.8kbit/s). 它有多种类型. 例如多脉冲激励线性预测编码器(MPE - LPC)、规则脉冲激励线性预测编码器(RPE - LPC)、码激励线性预测编码器(CELP)等. 多带激励线性预测编码器,也使用合成 - 分析法(AbS),改进了二元激励. 它能够在 2.4kbit/s 的速率下,得到较好的语音质量. 所有这些语音编码器都能够在单一 DSP(数字信号处理器)芯片上实现. 由于 DSP 芯片的运算能力不断增强,而价格又逐年降低,它不仅用于保密通信,而且广泛用于卫星通信、移动通信、短波通信和网络电话等很多方面.

6 语音的规则合成和文 - 语转换系统[3,7]

语音的规则合成是通过语音学规则产生语音的机器. 该系统内存储了较小的语音单位(如音素、双音素、半音节和音节)的声学参数,以及由音素组成音节,再由音节组成词和句子的各种规则. 当输入文字时,该系统利用规则自动地将它们转换为连续的语音. 目前,汉语合成技术大体上可以分为两类:

6.1 时域合成或称语音的波形合成

这种方案通常以音节为合成单位. 汉语共有 1280 多个单音节,可以从引导句中截取,经过适当的数据压缩后,构成一个汉语合成音节库. 使用时,根据要求的信息,从语音库中取出音节的波形数据,串接或编辑到一起,再经过重音、韵律、持续时间等修正,就可以输出连续的合成语音. 20 世纪 80 年代末,提出了基音同步叠加算法(PSOLA 算法),使得在波形数据的编辑过程中,能够方便地改变音调、重音、持续时间等物理特征,从而在组成词和句子时,能够方便地加入相应的规则,并转换为自然的、连续的语音. 这种语音合成技术,占用计算机的存储量较大,但合成语音清晰自然,目前使用比较广泛.

6.2 频域合成或语音的参量合成

仍以单音节、半音节为基本合成单元,首先从引导句中截取这些单音节、半音节的波形,并进行分析,计算出它们的物理特征参数. 主要的特征参数有 控制音强的幅度、控制音高(音调)的基频、控制音色的频谱(可以使用短时傅里叶变换或线性预测系数等). 线性预测系数也可以转换为共振峰频率和带宽,这样从语音学的观点考虑,更为直观. 这些参数经过编码压缩后,组成语音合成的参数库. 使用时,根据要求的信息,从参数库中取出相应的特征参数,经过编辑和连接,并加入语音合成所需要的规则,顺序送入到语音合成器. 在合成器里,这些参数控制着电子发声器官的相应部分,能够产生连续的语音. 这种合成技术所需要的存储器容量较小,但运算比较复杂. 为了改进合成语音的质量,也可以使用音调同步重叠相加的方法. 由于可以控制的参数比较多,而且和实验语音学联系紧密,也有很好的发展前景. 目前的语音质量正在不断地得到改善.

文 - 语转换系统是上述语音合成系统的进一步发展. 它输入的文字串是通常的文本字串,系统中的文本分析器根据发音词典,将输入的文字串分解为带有属性标记的词和相应的读音符号,再根据语义规则和语音规则,为每一个词、每一个音节确定重音等级、语句结构、语调、以及各种停顿等. 这样,文字串就转换为发出声音的代码串,合成系统就可以据此合成出具有抑、扬、顿、挫和不同语气的语句.

目前,这种系统已经被广泛地应用于社会生活的各个方面. 例如自动报时、自动报气象、电话咨询系统,以及用电话转发电子邮件等.

然而,人类的语音交流是涉及语言学、社会学、心理学、生理学等领域的复杂处理过程. 要使文 - 语转换系统能和播音员一样,具有情感并有很高的自然度,仍然是非常困难的问题. 它要求计算机对所朗读的文本,要有正确的理解. 这就要求计算机内要有一个丰富的知识库,还要有相当强的智能. 这是目前还没有解决的问题,有待今后深入研究.

7 语音识别和人 - 机语音对话[6—9]

语音识别包括发音人识别和语音识别两大类:发音人识别要从一群发音人中,找出预知他(她)声音的某一特定人. 它又分为与文本有关和与文本无关两种,前者要求发音人所说的文本是预先指定的,而后者要求文本是任意的和不受任何限制的,很显然,后者的难度更大.

语音识别有多种分类方法:按照词汇量的大小可划分为:小词汇语音识别(词数通常小于100);中等词汇语音识别(词数在100到500之间);大词汇语音识别(词数在500以上).目前已经做到好几万词汇.按照发音的方式,可分为孤立词语音识别和连续语音识别.孤立词识别是指发音者每次只说一个词或一条命令,它在词汇表中作为一个独立的识别单元,由识别系统来识别.连续语音识别是指发音人按照正常自然的发音方式发音,由识别系统来识别.按照服务对象可划分为:依赖于发音人和不依赖于发音人两种,即特定人工作方式和非特定人工作方式.凡识别系统只针对一个用户,即按照某一个特定发音人的特征而设计的,称为特定人工作方式.识别系统是根据很多发音人的共有特征设计的,允许任何人使用,则称为非特定人工作方式.

这些分类方法也可组合起来,形成多种语音识别系统.很显然,特定人、小词汇、孤立词语音识别系统是最简单的方式,比较容易实现.而非特定人、大词汇、连续语音识别则很复杂,虽然,目前国内有很多大学和研究所开发了可供表演的样机,美国IBM公司还推出了汉语连续语音识别软件.但是都还存在很多问题,没有得到推广和普及,未取得商业上的成功.

特定人、小词汇、孤立词语音识别系统大都采用简单的模板匹配原理.在训练阶段,用户将词汇表中的每一个词依次说一遍,并将它的特征矢量序列存入模板库中.识别时,将输入语音的特征矢量,依次与模板库中的每一个模板作相似度比较,相似度最高者就是识别的结果.但由于发音人在训练时和识别时,他们的说话速度不会完全一致,使得识别率难以提高,而使用动态时间伸缩算法(简称DTW算法),可以动态调整说话速度,从而找到最佳的模式匹配,使识别率提高.这种系统的识别率能达到98%以上,目前已经在一些自控装置、机器人等领域中应用.

非特定人、大词汇、连续语音识别系统的原理如图4所示.在预处理单元中,除了反混叠滤波器、模数转换器、自动增益控制外,还包括自动分段和识别基元选择.对于汉语,识别基元可用音素即声母-韵母,或者使用考虑了受前后发音影响的声母-韵母变体.一般地说,有限词汇量的识别基元应该选得大一些,而无限词汇量的识别基元应该选得小一些.声学参数可用倒谱系数,或者使用模拟人耳听觉特性的MEL谱,还需要加上能量、过零率、音调等特征.测度估计通常使用隐马尔柯夫模型(HMM).连续发音时,每一个音节甚至每一个音素都会受前后发音的影响,使得它的物理特征发生很大变化.再者,人们的发声器官都会有一些差异,不同发音人发出同一声音的物理特征,会有一些不同.这对于人的听觉器官来说,分辨语音信号的共性和个性,听懂和理解都能满意解决.但对计算机来说,却是很难的课题.目前最广泛使用的算法是隐马尔柯夫模型(HMM).马尔柯夫过程是一个双重的随机过程,人的语言过程就是这样一种双重随机过程.语言本身是一个可观察的随机序列,它是由大脑(不可观察的)根据语言需要和语法知识(状态选择)所发音素(或音节、词、句)的参数流.所以语音信号的模型可以用马尔柯夫模型来描述.马尔柯夫模型定义为

$$\lambda = F(A, B, \pi).$$

在这三个模型参数中,π 是事件(语音的参数流,可表示为矢量序列)的初始概率分布,B 是某状态下事件的概率分布,它就是外界观察到的事件符号的概率,A 是状态转移概率的分布.

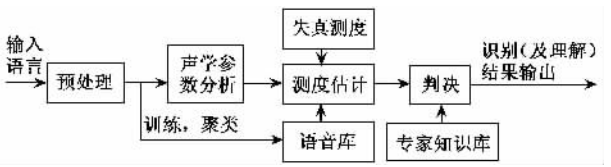

图4 非特定人、大词汇、连续语音识别系统的原理图

使用HMM作语音识别时,假设要识别的音素(或音节、词)有V个,为每一个音素(或音节、词)设计一个HMM模型.先用VQ技术设计一个尺寸为M(M为观察的符号数)的码本,然后用该音素(或音节、词)多次发音的语音数据,对它进行训练,得到最优的模型参数.与此同时,用最佳准则得到状态数为N的状态转移序列.最后,对实际要识别的语音信号用上面训练所得到的模型进行评估,吻合概率最大的那个音素(或音节、词)就是识别的结果.

8 结论和展望

本文简要介绍了“语音信号处理”这一分支学科的形成过程.并指出了它在现代信息科学中的地位.有一些基础的理论问题和技术问题还在继续研究和发展中.在信息科学的应用领域,例如语音的低速率编码,语音的规则合成和文-语转换系统,语音识别和人-机语音对话等,仍然是当前研究的热点.有的已经解决了,有的只是部分解决了,还有很多难

题等待我们去研究探索. 这些难题是：

(1)听觉器官的物理模型和数学表示,目前还没有一套权威的理论和成熟算法. 虽然有多种设计,但实验结果都不够充分. 特别是从听觉前端处理器所得到的波形特征,经过更高一级的处理,最后的信息速率只有 50 bit/s,这是人们理解和感知语音信号的最低限度. 而这一过程在人脑中是怎样完成的？目前还不太清楚. 它是一个非常复杂的问题,需要进一步研究探索.

(2)语音识别的子课题很多,其中最难的是非特定人、大词汇量、连续语音识别. 近年来这个课题已经取得很大的进展. 世界上有很多权威实验室推出了可供表演的识别系统,有些公司还推出了商品. 但是由于不同人的发音差别很大,再加上环境噪声等影响,系统的正确识别率和顽健性离实际使用还有很大距离. 目前,人们所期望的口呼打字机或听写机还没有得到推广.

(3)语音的规则合成和文－语转换系统,已经取得了一批可以实用的成果. 然而要使它能和优秀的播音员一样,具有不同风格、情感、很高的自然度,仍然是非常困难的问题. 关键技术是如何根据一段文章的语境和语义,自动生成计算机可以识别的韵律符号. 这涉及到机器对自然语音的理解,目前还在研究中.

(4)语音增强包括从强噪声中提取语音信号,或者从几个人同时说话的混合波形中,分离出各自的语音信号,这类研究虽然理论上有一些算法,但效果均不理想,还没有达到可以实用的水平.

(5)最后谈一下大家感兴趣的课题——语音翻译机. 如果前面所说的非特定人、大词汇量、连续语音识别、机器对自然语音的理解和处理、语音的规则合成和文－语转换系统等课题,都满意地解决了,则输入为英语、输出为汉语的语音翻译机(或者相反、或为其他语种)也就应运而生了. 这将会对旅游、商务和文化交流带来深远的影响. 早在 10 年以前,美国、德国和日本三家合作,已经在研究这样的机器了. 而且已经有词汇量不大、内容单一、可供表演的机器. 当年江泽民主席访美,在中美科技合作计划中,也有这样的项目. 然而,这毕竟是一个很难的课题,它涉及广泛的科学技术问题,目前仍在研究和开发的过程中.

参 考 文 献

[1] Kondoz A M. Digital speech coding for low bit rate communication system, University of Surrey. UK, John Wiley & Sons, Inc 1994

[2] 李昌立,吴善培编著. 数字语音——语音编码实用教程. 北京:人民邮电出版社,2004. 11 [Li C L, Wu S P. Digital speech——Practical Course of Speech Coding Beijing: Post & Telecom Press 2004. 11(in Chinese)]

[3] Thomas F. Quatieri Discrete－time speech signal processing, Principles and Practice, Massachusetts Institute of Technology, Lincoln Laboratory, Prentice Hall PTR, 2001

[4] 马大猷著. 语言信息和语言通信. 北京:知识出版社,1983 [Ma D Y. Language Information and Language Communication Beijing: Publishing House of Information 1983(in Chinese)]

[5] Lajos H F, Clare A S, Jason P W. Voice compression and communication, Principles and application for fixed and wireless channels. John Wiley & Sons, Inc, 2001

[6] 杨行峻,迟惠生等编著. 语音信号数字处理. 北京:电子工业出版社,1995 [Yang X J, Chi H S. Digital Processing of Speech Signals Beijing: Publishing House of Electronics Industry 1995(in Chinese)]

[7] Kleijn W B, Paliwal K K. Speech coding and synthesis. Elsevier Science B. V, 1998

[8] Jean－Luc Gauvain, Lori Lamel, Large－Vocabulary continuous speech recognition: Advances and Applications Proceedings of the IEEE, August 2000

[9] 陈永彬,王仁华. 语音信号处理. 合肥:中国科学技术大学出版社,1990 [Chen Y B, Wang R H. Processing of Speech Signals Hefei: Publishing House of University of Science and Technology of China 1990(in Chinese)]

·物理新闻和动态·

豌豆大小的磁力计

一种只有豌豆般大小的磁力计已被美国国家标准与技术研究所(NIST)的 P. Schwindt 博士领导的研究组研制成功. 这种磁力计的基本原理是利用铷原子的各个量子能级的能量大小是与其周围的磁场有关的这一事实. 他们将一小部份的铷原子压缩在一个胶囊内,并让激光通过原子形成一个可读的记录用来作精确的测量. 整个装置的大小只有 $12mm^3$. 但是这个小磁力计可以在许多方面对磁场进行测量,且其测量的灵敏度可达到 50pT. 另一方面,这种磁力计可以藉助于平板印刷技术进行大批量的生产,因此其成本很低. 这个装置的另一优点是它要消耗的能量极少. 显然这个装置可在地球物理与地质探测方面得到广泛地应用,特别是对地下与水下的各种铁器作有效的探测,例如探测地下的铁管、坦克、沉入海中的海船以及其他物品等. 总之,这种小体积、低成本的磁力计要比目前存在的磁力计的设计在各方面要先进很多.

(云中客　摘自 Applied Physics Letters, 27 December 2004)