语音总体效能的计算机客观评价方法研究*

陈越 贺汉根

(国防科技大学自动控制系 长沙 410073)

摘 要 多媒体界面中的语音评价问题尚处于研究阶段。本文探讨了语音评价的基本方法,重点对语音的客观评价方法进行研究,提出了一种应用语音分析技术进行语音总体效能客观评价的计算机处理方法。

关键词 人机交互,效能分析,语音评价,语音分析

分类号 TN912.3

Research on Computer Processing Method to Evaluate the Overall Effectiveness of Speech

Chen Yue He Hangen

(Department of Automatic Control, NUDT, Changsha, 410073)

Abstract The evaluation method of speech on multimedia interface is being studied. This paper discusses Some basic methods of speech evaluation, putting the emphasis on the research of the objective evaluation of speech. Using analytical technology of speech signal, the authors have developed a computer objective evaluation method to evaluate the overall effectiveness of speech.

Key words human-computer interface, effectiveness evaluation, speech analysis, speech evaluation

多媒体界面已成为人机界面发展的主流,而多媒体界面的效能分析则是一个很有意义、也很复杂的研究课题,国内外许多学者都已开展这方面的研究工作,而且取得了一定的进展[1~3]。本文讨论的是多媒体界面中的语音评价问题。

文章对语音评价的基本方法进行了探讨,重点研究语音的客观评价方法,提出了一种应用语音分析技术、由计算机对语音总体效能进行客观评价的处理方法。

1 语音评价的基本方法

在人的感觉系统中,听觉是仅次于视觉的重要感觉。对语音信号作评价,涉及到语音感知问题,这和心理声学、心理语言学、生理语言学、发音语言学等学科都有密切联系,因此语音评价是一个复杂的问题。

语音评价可以是主观评价,也可以是客观评价。可以用于效能分析的技术也很多,例如分析方法、实验方法、仿真技术等等。在不同的分析过程中,所应用的分析技术也各不相同。

1.1 主观评价方法

由于客观测试结果还不能完全反映主观的感觉,因此听觉评价常采用主观评价方法,即采用会面、调查、询问等方法,提交人的操作报告。具体过程包括:

(1) 准备恰当的主观调查表;

(2) 通过被调查人群对交互系统的实际操作,对调查表的各项内容评分;

* 国防预研基金项目资助
1998年4月2日收稿
第一作者:陈越,女,1968年生,讲师

(3) 利用统计技术, 对收集到的主观数据进行定量和定性分析, 得到主观评价的结果。

主观评价的结果是否公正、客观, 关键在于调查表的准备和调查人群的确定。调查表中的项目要从心理声学的角度, 依据语音评价条目确定。另外应选择具有正常听力的调查人群, 人群以专业人员为主, 非专业人员为辅。即调查对象要挑选得能对预期的用户团体具有代表性, 并注意到他们在计算机的使用、任务的经验、所受教育的程度、年龄、性别等方面的背景。只有选择的调查人群数量足够, 代表性齐全, 才能保证主观评定结果的客观、公正、可靠。

1.2 客观评价方法

声音虽然是客观存在的, 但是人类的主观感觉(听觉) 和客观实际(声音) 有一致的地方, 也有不十分一致的地方, 甚至还会产生 "错觉 "。而且随着人们种族、年龄、性别、生活经历、听音素养等许多因素的不同, 他们的主观感觉(包括对语言的理解程度) 有很大的差别, 因此往往使得主观评价结果比较分散。这就需要有一种客观评价方法, 它不受主观因素的影响, 而以听觉的物理特性为标准, 即评价基于语音的客观性能指标进行。

我们知道, 影响主观听觉效果的主要物理量是语音信号的频率和声强[4]。进行客观评价最直接的方法, 便是提取多媒体界面中语音信号的频率、声强等物理量的精确值, 代入评价准则, 立即得到评价结果。但是要这样做, 第一, 必须有能精确采集频率、声强等物理量的实验设备; 第二, 要有非常细致的评价准则, 使具体到各种频率、声强值的组合, 都有明确的评价, 而这应以大量的心理、生理实验和数据统计为依据。

在研究中, 我们提出了一种语音总体效能的计算机客观评价方法。该方法分两步: (1) 利用语音分析技术提取相关的物理量; (2) 利用提取的物理量, 给出语音的总体效能评价。下面做详细介绍。

2 语音信号的分析和相关物理量的提取

我们通过对语音信号作短时分析[5]提取相关物理量。

进行听觉评价, 首先必须获取语音信号。受实验条件所限, 我们对存放成语音文件(.WAV) 格式的语音信号进行分析。在研究中, 我们主要利用短时傅里叶分析获取语音信号的平均频率, 利用短时平均能量判断语音的起止点, 确定语音的平均能量, 并得到语音的语速。

(1) 平均频率 F_{ave}的获取

若有语音离散信号 $x(m)$, 语音信号的离散短时傅里叶分析可以表述为:

$$X_n[e^{j\frac{2\pi k}{N}}] = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{-j\frac{2\pi km}{N}} \qquad 0 \le k \le N-1 \tag{1}$$

这里窗函数选用 Hamming 窗, 即

$$w(n) = \begin{cases} 0.54 - 0.64\cos(2\pi n/(N-1)), 0 \le n \le N-1 \\ 0 \qquad\qquad\qquad\qquad \text{else} \end{cases} \tag{2}$$

此处窗长取 $N=128$(或 256) ,

利用 FFT 式(1) 计算, 可以得到该语音信号的频率。将每帧的频率取平均, 即得到该语音信号的平均频率 F_{ave}。而每帧频率的不同, 则反映语音信号的频率变化。

(2) 平均能量 E_{ave}的获取

语音信号的短时能量函数可定义为:

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m) \tag{3}$$

此处窗函数仍取 Hamming 窗, 定义如(2) 式, 窗长取为 128。

将每帧的短时能量取平均, 即得到该语音信号的平均能量 E_{ave}, 而每帧能量的不同, 反映了语音信号的幅度变化。

要说明一点, 如此得出的能量, 和人耳所感受到的响度, 以及通常所说的声强有所区别。响度以及声强, 和语音实际播放时的环境、器件等因素有关, 要获得准确的值, 需要有特定的仪器。为简单起见, 我们

用短时能量的方法来判别相同环境条件下语音信号的相对大小。

(3) 平均语速 V_{ave}的计算

语言传播的一个重要评价指标是语言清晰度,这一般是靠实验测试得到的。我们认为,人在一般情况下的语速应该是比较平稳的,而且每个人说话的快慢对语言清晰度也有一定的影响,为了得到尽可能多的客观物理量来作评价,我们进行了平均语速的计算。

计算平均语速的关键是在输入的语音信号中,准确地找出语音段的起止点。鉴别语音端点的主要特征是能量。我们通过做短时能量分析来完成端点判断,确定语音数据的长度。然后根据语音信号的采样频率,求出语音段的时间,将之除以语音字数,即可得到平均语速。具体步骤如下:

(a) 利用式(3) 计算所有帧的语音能量;

(b) 求最大能量值 E_{max};

(c) 确定能量的门限值 $ET = E_{max} \times 2\%$;

(d) 利用能量门限值确定端点,能量第一次超过门限的点为语音的起点 S,能量开始低于门限的点为语音的终点 D。

(e) 根据语音的起止端点,求出语音数据的长度 $Length = D - S$;

(f) 根据语音信号的采样频率 F_s,求出语音段的时间 $T_s = Length / F_s$;

(g) 根据已知的语音字数 N,求得平均每个字发音所需的时间 $T_{ave} = T_s / N$;

(h) 计算平均语速 $V_{ave} = 1 / T_{ave}$。

在用短时能量方法进行端点判断时,要采用一定的保护措施。首先,门限值要根据语音的一般强度进行归一化处理。如果语音信号很弱,则门限值也要取低。我们在实验时,如(b)、(c)两步所示,是采用最大能量值的一定百分比作为门限值的。当然,门限值的确定,需通过多次实验。另外,还必须有防虚警措施。在没有语音时,随机噪声也可能在一瞬间超这门限值,这不应引起识别器的误动作。因此对信号还要有时间要求。例如,在识别时,我们在发现能量强度超过门限的时间已有5帧时,才反过来规定第一帧为语音的起点。对终点判断也采用类似的方法。当然,也可以规定两个门限值,并且当能量超过高门限时,再以超过低门限的时间为语音的起点。

平均语速的计算精度主要受端点检测的影响。端点检测技术在高信噪比的实验室条件下容易实现,而在实际应用中很难保证完全准确。另外当语音段的首尾是能量很低的清音时,端点检测也容易发生错误。

(4) 识别误差率 E_r

我们尝试了通过对语言信号进行单词划分来判断语言清晰度的方法。该方法出于这样的考虑:假如每个字发音很清楚,则我们认为字与字之间应是能够划分的。具体地,即利用语音信号的短时能量函数,通过选择恰当的能量阈值,进行单词划分,得到每句话里的单字个数 n,同时交互输入实际的语音字数 N。我们定义识别误差率为:$E_r = |n - N| / N$。单词划分的关键是能量阈值的确定,这需通过多次实验。因为不同的音素之间其实没有明显的分界,几乎每个音素都逐渐消失在其后面的音素中,另外不同的音调发音也不同,因此普遍适用的阈值很难确定。在实验时,我们把单词的划分结合到语音端点的判断中。

单词划分只反映了人说话时吐字清晰与否,而忽略了人在听的时候所具有的理解力和文字的上下文相关性。

3 语音总体效能的客观评价

我们重点研究了如何利用统计方法和相关技术给出主观评价和客观物理量之间的关系,并由此得到语音的总体评价。

总体评价应该是对各物理量的加权平均,关键问题是各个权值的确定,即寻找使加权平均结果与实际主观评价的统计结果最吻合的一组最佳权值,这可以视为一个学习的过程。和主观评价一样,选择恰当的调查人群非常重要,另外学习时所选用的语音样本也很有讲究,应具有相当的代表性。我们采用的具体方法如下:

首先设计主观调查表, 请所有的被调查人员对语音对象 i 的总体满意程度作出评价(十分制), 求出总体满意度均值 S_i。设 $W_1, W_2, \cdots, W_p$ 分别是各个物理量相应的权值, 另外对该语音对象作语音分析所提取出来的 p 个物理量为 $X_{i1}, X_{i2}, \cdots, X_{ip}$。若有 n 个语音对象, 则得到 n 组统计数据($S_i, X_{i1}, X_{i2}, \cdots, X_{ip}$)($i = 1, 2, \cdots, n$)。我们利用最小二乘原理, 采用下述线性表达式:

$$S_i = W_1 * X_{i1} + W_2 * X_{i2} + \cdots + W_p * X_{ip} + W_0 \qquad i = 1, 2, \cdots, n \tag{4}$$

对这 n 组统计数据进行多元线性回归, 则最后确定的权值($W_1, W_2, \cdots, W_p$)即为与实际主观评价的统计结果最吻合的一组最佳权值。

实验中我们得到的一组最佳权值如表 1 所示。

表 1

物理量	平均频率	平均能量	平均语速	识别误差率	常数
系数	- 16.5	0.45	- 0.26	- 0.53	0.943

有了最佳权值以后, 进行客观评价时, 只要通过语音分析提取到相关的物理量, 再进行加权平均, 就可以得到对该语音信号的总体评价(评分)。

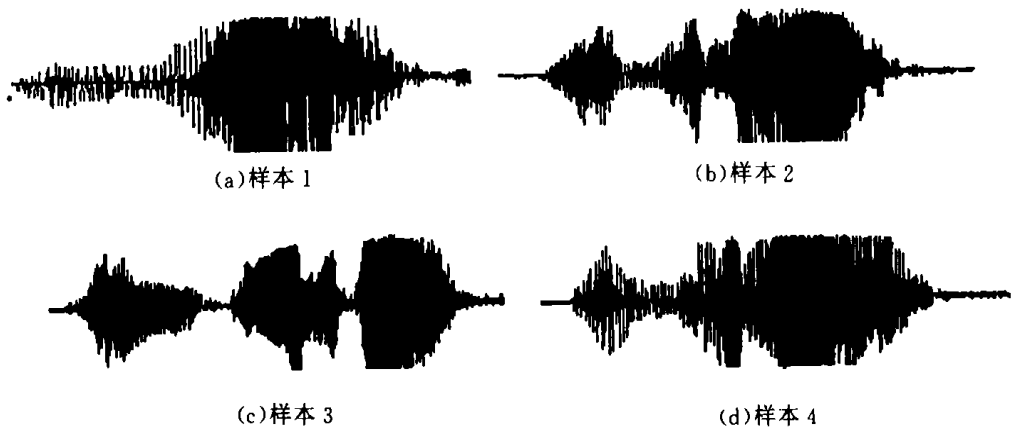

图 1 样本语音波形图

实验中, 我们重点对告警语句进行了研究和分析。图 1 所示的四幅语音波形, 说的是同一句话: "有故障! '。从波形可以看出, 这四句话的语音质量很不相同, 表 2 给出了计算机所得到的总体评价和人的主观评价的比较。

表 2 计算机总体评价和主观评价的比较

	样本 1	样本 2	样本 3	样本 4
计算机总体评价分数	3.23	5.58	8.27	6.16
主观评价	不好	还可以	很好	一般

从表 2 可以看出, 由计算机所做的客观评价, 与人的主观评价基本是一致的。这也证明了该方法的可行性。

在实验时, 我们还发现, 如果是作为学习样本已经被训练过的语句, 那么即使说话者是任意的, 其评价结果与主观感觉相差不大。相反, 若是要评价的语音对象并未被学习过, 则评价结果可能会与主观感觉产生较大出入。分析原因, 主要是由于语音信号涉及到发音学、语言学, 说不同的话时, 不同的发音方法导致了语音频率、能量之间的较大差异。故而该方法对经过训练的语句是可行的, 普遍适用性要差一些。在具体运用时, 针对评价对象, 要有一个预学习的过程。幸而一般而言, 语音评价应在系统设计的过程中进行, 做到这一点并不难。

4 结束语

语音效能分析尚处于研究、发展阶段，至今尚未形成完整的理论和方法。总的看来，进行语音评价比较困难。要合理地做到客观评价，要求具备相当准确的物理量，但这点又很难做到。例如，我们曾希望能对语音信号的警觉程度作出客观评价，但很难找到一个物理量与之直接相关联。事实上，语音信号本身既包含与说话人有关的量，也包含有语言信息，而这两者一般是不易区分的，而且警觉感还和个体有关，不同的人对言语紧张的理解就不一样。单就语音效能分析的意义上说，应用统计处理方法，同时结合语音分析技术，给出客观评价应是一个实用的可行途径。

参考文献

1 李成辉，谭浩，刘锦德. 自适应人机界面的评价. 计算机科学，1995，22(4)：69～74

2 Malone B T，Kirkpatrick M，Heasly C C. Human-Computer Interface Effectiveness Evaluation. In: Gavriel Salvendy. Human-Computer Interaction. Amsterdam: Elsevier Science Publishers B. V.，1984: 117～120

3 Norman D A. Cognitive Engineering Principles in the Design of Human-Computer Interfaces. In: Gavriel Salvendy. Human-Computer Interaction. Amsterdam: Elsevier Science Publishers B. V.，1984: 11～16

4 丁玉兰. 人机工程学. 北京: 北京理工大学出版社，1991

5 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