

基于特征比较的语音评分方法研究

刘振安, 罗永钊

(中国科学技术大学 自动化系, 安徽 合肥 230027)

(hkeelyz@mail.ustc.edu.cn)

摘要:为在语言朗读训练时对跟读语音的质量自动作出客观评价并在嵌入式系统上实现,提出了一种基于特征比较的语音评分方法。通过分析输入语音,提取语音特征并与参考标准进行匹配比较,由评分机制根据相似程度大小给出评价得分。实验证明该方法的算法复杂度较低,评分结果符合人的主观感觉。

关键词:语音评分;动态时间规整法;基频轨迹;梅尔倒频谱参数

中图分类号: TN912.34 **文献标识码:** A

Research of speech assessment based on feature comparison

LIU Zhen-an, LUO Yong-zhao

(Department of Automation, University of Science and Technology of China, Hefei Anhui 230027)

Abstract: In order to impersonally assess the quality of the following speech in the language speaking training, and carry out it in embedded system, a method of speech assessment based on feature comparison was proposed. The input speech was analysed, and the feature was extracted to match with the reference speech. The result was given by the mechanism based on the similarity of the speech. Experiment shows that the algorithm complexity is low, and the result of the assessment is approximately consistent with the subjective feeling.

Key words: speech assessment; dynamic time warping; pitch contour; mel-frequency cepstral coefficients

1 语音评分的系统构成

在外语学习中,语音评分能作为学习过程中的交互手段,给学习者提供辅助性的指导,并提高语言朗读的能力。语音评分普遍使用音素评价方法。它是基于语音模型的评价方式,通过语音识别技术切割出每个音素单元,再对每个音素单元和模板库里的音素进行比较并求得相似度^[1,2,5]。这种方法的评分结果只反映了语音内容的正确程度,却忽视了语言朗读过程中的韵律变化特点,不仅需要对大量样本数据进行训练以建立语音模型,还需要借助较强运算能力的平台(如PC机)实现语音识别功能,运算量较大,不适合嵌入式系统。

本文研究的语音评分是一种基于语音特征比较的评价方法。它通过对比参考标准语音与待评价语音,从一个比较主观的角度去评价一段语音的质量。文献[2,5]指出,梅尔倒频谱参数和音调参数对语音评分的重要性较大,因此本文讨论的语音评分方法把评价的重点放在音调韵律变化的比较上。这样做可减少评分参数的数量,并在保证评分质量的基础上降低评分操作的运算量,使得评分系统能够从PC平台移植到嵌入式系统上。

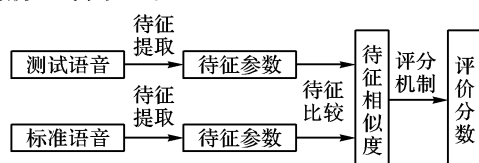


图1 系统流程

语音评分可分为三步:(1)对输入的语音提取特征参数;

(2)利用模式对比的方式对两者的特征参数进行比较;(3)评分机制根据特征相似度对语音作出评价。图1是评分系统的流程图。

2 评分算法的实现

2.1 特征参数的提取

语音信号是一种时变信号,但可以利用语音信号的短时平稳性,在一小段时间片上对信号进行观察,及时跟踪信号的变化。本文使用基频变化轨迹(Pitch Contour)和梅尔倒频谱参数(Mel-Frequency Cepstral Coefficients)来对输入语音进行评价。梅尔倒频谱参数表征声音的内容。而基频轨迹则显示声音音高的变化,反映音调的起伏与韵律的变化。

2.1.1 基频参数的提取

基音周期采用中心削波的AMDF(Average Magnitude Difference Function)^[5]求取。AMDF相比于常用的基于相关、基于同态信号处理、基于线性预测编码的基音估计器,它不涉及乘法和除法,比较适合应用在嵌入式系统等平台上;在基音周期点的谷点要比短时自相关函数的峰点尖锐,因此估计精度较高,较稳定。对于第 n 帧语音信号 S_n :

$$AMDF_n(t) = \frac{1}{M} \sum_{m=0}^{M-t-1} |S_n(m) - S_n(m+t)| \quad (1)$$

M 为帧长, t 为候选基音周期。当 $S_n(m)$ 是周期性的,周期为 T ,则 $AMDF_n(t)$ 在基音周期 T 或其整数倍点附近包含谷值。因此可选取第一个谷值点作为对该段语音的基音周期。估计的基频为:

收稿日期:2005-06-24;修订日期:2005-09-05 基金项目:国家自然科学基金资助项目(60272040)

作者简介:刘振安(1947-),男,江苏徐州人,教授,主要研究方向:语音压缩和模式识别; 罗永钊(1980-),男,广东佛山人,硕士研究生,主要研究方向:语音处理与嵌入式系统应用。

$$F_0(n) = \frac{1}{\min_{t>0}(\arg \text{localMin}(AMDF_n(t)))} \quad (2)$$

LocalMin 为求取局部最小值的函数。

实际计算中,第一最小谷值点的位置有时与实际的基音周期不吻合,主要是受到声道的共振峰特性造成的干扰。因此,在用 $AMDF$ 进行基音周期估计前,要对语音信号进行预处理消除共振峰的影响,改善估计效果。预处理的方法使用中心削波非线性变换。经削波变换的信号通过 $AMDF$ 后在基音周期点上的波谷变得更明显,基音周期估计的效果可得到一定的改善。另外,还要对求得的基频曲线进行线性平滑处理,以消除误差较大的野点。

获取的基频参数是语音评分中的一个重要的参数,它的高低变化反映了语音的语调韵律特点。模仿标准读音的语调在语言朗读学习中是一个基本的要求。利用音调变化进行评分是该评分算法的一个特点,它有别于以往的基于内容准确性的语音评分算法,从一个较高的朗读层次去评价一段语音的质量。

2.1.2 梅尔倒频谱参数的提取

语音信号经过预加重处理后,语音信号受到压抑的高频部分得到补偿。为减小 Gibbs 效应,对每一帧语音信号加上一个汉明窗 (Hamming Window)。经过快速傅立叶变换 (FFT) 后可获得每一帧的频谱,通过一组 20 个三角滤波器得到每个频带的输出对数频谱 $m_j (j = 1, 2, \dots, 20)$, 经过离散余弦变换 (DCT) 即可求得 12 维的梅尔倒频谱参数:

$$c_j = \sum_{k=1}^p m_j \cos\left(\frac{\pi k}{p}(j - 0.5)\right), k = 1, 2, \dots, 12 \quad (3)$$

其中 $p = 20$, 为三角带通滤波器的数目。

2.1.3 特征参数的规整化

由于语音的基音频率刻画了说话人的声带特性,因此存在一些个人的音高差异。本文研究的语音评分着重的是声调韵律的变化,为了更好地对不同的语音进行比较,需要对特征参数进行规整化处理,使得存在个体差异的特征参数可以在同一基准下进行比较。

2.2 模式对比

为了比较待评分语音与参考标准语音,可以通过估测两者的特征参数的差距来反映它们之间的相似度。由于两者在朗读语速、停顿时间等方面均不相同,不能够对两者直接进行比较,因此这里采用动态时间规整法 (Dynamic Time Warping) 来求取两者之间最相近的比较路径。一般的算法都是对各个特征参数分别进行 DTW 校正,这样做不仅增加算法的运算量,而且割离了特征参数之间的相互关联性。本文的模式对比算法则采用一种综合的方式改善了这种不足。它首先利用 MFCC 参数来对两段语音进行 DTW 非线性校正,使得输入语音和标准语音在内容相似的位置是相互对应的,这跟音素评分方法中切割出语音音素的目的是相同的。此时可以得到一条误差最小的校正路径 path 和对应的 DTW 距离,该距离是两段语音的 MFCC 特征的比较结果,反映了两段语音在内容上的发音差别。基于这条校正路径,基频变化轨迹 $f_1(n)$ 和 $f_2(m)$ (分别对应待评分语音和参考标准语音) 即可在相似内容的对应位置进行比较。比较的对象是基频点的差距 $|f_1(n) - f_2(m)|$ 与其变化量的差距 $|\Delta f_1(n) - \Delta f_2(m)|$ 。其中, $\Delta f_1(n) = |f_1(n) - f_1(n-1)|$ 。差值越小,表明两者语调越相似。由此可见,只需经过一次 DTW 操作,就可以结合两种特征参数进行模式对比。

假设参考标准语音的 MFCC 特征向量为 $M_1 = [m_1(1), m_1(2), \dots, m_1(T)]$, 基频特征向量为 $P_1 = [p_1(1), p_1(2), \dots, p_1(T)]$ (T 为参考语音的长度); 待评价语音的 MFCC 特征向量为 $M_2 = [m_2(1), m_2(2), \dots, m_2(S)]$, 基频特征向量为 $P_2 = [p_2(1), p_2(2), \dots, p_2(S)]$ (S 为待评价语音的长度)。则:

$$C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \text{DTW}(M_1, M_2) \quad (4)$$

$$\begin{pmatrix} P \\ M \end{pmatrix} = \begin{pmatrix} P_1 & P_2 \\ M_1 & M_2 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \quad (5)$$

其中 C 是特征比较矩阵,它是利用语音的 MFCC 特征向量进行 DTW 后而得到的。利用该特征比较矩阵,可以求得基频变化相似度 P 和 MFCC 特征相似度 M 。

2.3 评分机制

语音评分的目的是为了显示这段声音的发音是否正确规范,语调是否符合要求。分数越高,则表明对这段声音的满意度越高,反之,低分数表明这段声音的发音不够准确或没达到满意的要求。文献[4,5]中提出的评分机制是一种评分参数相互独立的加权组合,没有考虑到评分参数之间的相互联系性。本文的评分机制则把这种关联性作为评分的一部分,从一个比较全面的角度去衡量语音的朗读质量。评价分数可定义为:

$$\text{scores}(P, M) = k_1 P + k_2 M + k_3 PM \quad (6)$$

其中 k_1, k_2, k_3 为各评分参数在评分中的权值, P 为基频变化相似度, M 为 MFCC 特征相似度。权值的选择可以根据不同的要求或评分的侧重点不同而有所不同。为了使计算机能够更好地模拟语言专家的评分,可以对权值进行训练,找出计算机评分和人工评分的一个最佳映射关系。

3 结果分析

在测试中,选取一个测试语句: Would you say that again? 测试者第一次说的是完整的句子,如图 2 所示。第二次则少说 that 这个单词,即: Would you say again? 如图 3 所示。评分采用的是百分制。评分结果,第一段语音的得分是 60.44,第二段语音只得到 20.71 分。可见评分系统对句子内容的完整性比较敏感。

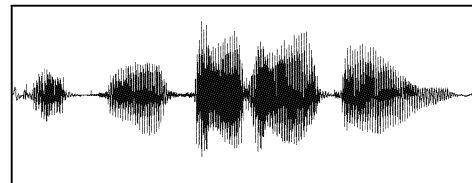


图2 完整的句子

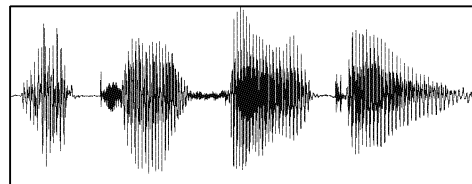


图3 缺少单词“that”的句子

表 1 为 8 位测试者分别读 10 个句子的评分结果。在评分过程中,如果输入的语音是一段清音或者背景噪音,则评分不会正常进行,这是评分算法中判断干扰噪音的结果,这样可有效排除非正常输入语音,提高评分效率。评分系统对语音的音量大小不敏感。若把测试语音的音量增大一倍,评分结果只下降 0.2%。

表1 8位测试者分别读10个句子的评分结果表

	1	2	3	4	5	6	7	8	9	10
A	60.55	62.22	73.07	67.45	66.08	61.88	59.70	55.83	78.22	74.11
B	33.73	57.52	55.90	60.38	65.67	55.92	32.31	48.88	56.08	56.06
C	51.87	78.27	63.24	75.58	80.05	73.08	53.31	77.25	71.19	74.69
D	49.82	24.02	52.77	53.48	60.37	59.93	47.27	39.60	45.70	58.64
E	54.13	76.94	57.72	57.30	72.35	60.61	60.74	58.02	68.48	60.50
F	26.65	42.58	53.08	67.37	67.48	47.30	41.27	67.75	69.75	62.93
G	54.83	59.89	53.18	67.13	67.11	53.28	42.73	62.11	60.25	67.74
H	62.28	81.95	52.15	65.67	77.51	59.91	43.28	51.32	68.96	71.00

评分结果作为一种反馈信息,反映出输入的朗读语音与参考标准语音在语调韵律上的相似性,该结果比较符合人的主观感觉。由于评分是基于参考标准语音的比较,为了提高评分的质量,需要把朗读质量较好的语音作为参考标准语音。

与音素评分算法相比,本文算法的特点是不仅对语音内容的准确性作出评价,而且还考虑了朗读过程中的韵律变化因素,从整体的角度去评价句子朗读质量的好坏。因为不需要繁重的语音模型训练运算和预存模型特征库,所以节省了

存储空间和运行时间。这对在嵌入式系统上实现评分功能是一种可行的解决方案。

参考文献:

- [1] NEUMEYER L, FRANCO H, WEINTRAUB M, *et al.* Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech [A]. Fourth International Conference on Spoken Language Proceedings, ICSLP 96 [C]. 1996, vol 3. 1457-1460.
- [2] YORAM M, HIROSE K. Language training system utilizing speech modification [A]. Fourth International Conference on Spoken Language Proceedings, ICSLP 96 [C]. 1996, vol 3. 1449-1452.
- [3] THOMAS F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice [M]. 北京:电子工业出版社,2004.
- [4] CHEN JC, LO JL, JANG JSR. Computer assisted spoken English learning for Chinese in Taiwan [A]. International Symposium on Chinese Spoken Language Processing [C]. 2004. 337-340.
- [5] 李俊毅. 语音评分 [D]. 台湾清华大学硕士论文,2002.

(上接第2920页)

在局域网环境中建立测试系统原型,其中一台作为中央服务器运行信息服务代理,在中央服务器上定义虚拟资源的目录,动态调整每种资源的特殊访问代价 U , 并且动态定义这些资源的价格,即在某个范围内随机变化资源的价格;在其他各个客户端上运行应用程序,随机访问虚拟资源,并计算访问所花费的总代价;在各个客户端上运行存储资源代理程序,将本地磁盘作为 cache,根据 cache 中缓存资源的使用频度预测其价值,以 $e^{-C/\Delta V}$ 的值与阈值 P_0 来决定是否缓存新资源,并统计本地访问资源的命中率。

在该系统上运用最近最久未使用 cache 替换策略,分别使用传统的 cache 接受策略与本文的基于经济的优化策略,运行5次,每次持续运行24小时,则两种情况下5次访问的命中率如表1所示。

表1 不同接受策略下的 cache 命中率比较

	不同访问次数下的命中率 (%)				
	1	2	3	4	5
non-elec	36	34	38	37	39
elec	34	35	33	34	36

由表1可知,平均命中率的变化不明显,相比之下,客户端访问资源的花费却有相当大的变化,5次访问的平均花费比较如表2所示。

表2 不同接受策略下的资源访问花费比较

	不同访问次数的平均花费/元				
	1	2	3	4	5
non-elec	135.1	127.4	131.3	129.7	131.8
elec	97.7	103.2	102.6	104.0	99.8

由模拟结果可知,对网格 cache 接受策略进行基于经济的优化,可以在平均命中率无明显变化的情况下,使访问代价大幅度的降低,这对于对价格敏感的用户非常有利。

本文所讨论的网格 cache 的接受策略,根据市场中的经济理论对所使用资源的价值进行评估,其价值主要是根据资源的价格和使用频度的综合考虑,在资源的预测价值基础之

上尽量减小访问延迟。该策略只是涉及到简单的计算,其本身的成本较低。

4 结语

对网格 cache 接受策略进行了讨论,并引入了经济模型,其目标是在访问延迟无明显降低的情况下,使本地代理以尽量小的代价完成所需要的资源请求。首先对存储资源代理进行简单阐释,进而定义如何对资源对象的价值进行评价,最后说明本地存储资源代理对应用所访问的资源进行价值评价,并确定是否缓存的网格 cache 接受策略。

本文的网格 cache 接受策略,主要是从用户的利益角度出发,使用户执行网格任务所花费的代价尽量小,是在使用 cache 的收益角度来提高 cache 的命中率,而不仅仅是从 cache 的内容方面。

参考文献:

- [1] FOSTER I, KESSELMAN C, TUECKE S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations [J]. International Journal of Supercomputer Applications, 2001, 15(3): 1-10.
- [2] 陈梅, 都志辉. 网格 Cache 若干问题分析 [J]. 计算机科学, 2004, 31(5): 15-17.
- [3] BELL WH, GAMERON DG, CARVAJAL-SCHIAFFINO R, *et al.* Evaluation of an Economy-Based File Replication Strategy for a Data Grid [A]. Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'03) [C]. Tokyo, Japan. 2003.
- [4] CARMAN M, ZINI F, SERAFINI L, *et al.* Towards an Economy-Based Optimisation of File Access and Replication on a Data Grid [A]. Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02) [C]. Berlin, Germany. 2002.
- [5] OTOO E, SHOSHANI A. Accurate Modeling of Cache Replacement Policies in a Data Grid [A]. Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSS'03) [C]. San Diego, California. 2003.
- [6] 曹鸿强, 肖依, 卢锡城, 等. 一种基于市场机制的计算网格资源分配方法 [J]. 计算机研究与发展, 2002, 39(8): 913-916.