文章编号: 1671-4229(2011) 01-0065-05

语音质量客观评价方法研究及实现

方凡泉，李心广，王桂珍，林江豪*
(广东外语外贸大学,广东 广州　510006)

摘　要: 语音质量的客观评价在语言自主学习中具有重大的意义. 文章首先介绍了语音质量客观评价过程中语音信号的预处理单元; 其次,介绍了不同的语音特征提取算法,比较选择了更符合人耳听觉模型的 MFCC 特征,并给出特征提取过程及结果; 最后,比较当前评价模型(DTW 和 HMM) 的优缺点,并提出了采用 HMM 模型进行评价的方法,设计系统验证了该方法下评价的客观性.

关键词: 语音质量; 客观评价; HMM 模型; 语音特征
中图分类号: H 01　　**文献标志码**: A

语音质量的评价主要是对语音的音准、语调等指标做综合评价. 对于语音音素、单词,在评价过程中主要以音准为主要指标. 评价模式可分为主观评价和客观评价,主观评价主要是人工对语音质量进行逐项评价; 客观评价是利用计算语言,实现机器智能化的评价. 在进行英语语音考试评分中,主观评价工作量非常大[1]. 目前,语音发音质量的自动评价方法比较多[2-5],但均存在评价结果不客观的缺点. 因此,本文采用基于人耳听觉模型的 MFCC(Mel-Frequency Cepstral Coefficient) 特征作为评价特征; 提出建立 HMM(Hidden Markov Model) 模型进行特征匹对,实现语音质量的客观评价. 最后,对 10 位英文专业的学生进行录音,录制 5 个音素及其对应单词,并将录音用于系统评测. 经检验,系统有比较客观的评价结果.

1　评价过程

本文采用如图 1 所示语音评价过程. 首先,对语音信号进行采集,应用 PC 的声卡,将模拟信号数字化,根据奈奎斯特(Nyquist) 采样定理: 在进行模拟/数字信号的转换过程中,采样信号的频率最低等于被采样信号频率的 2 倍,则采样之后的数字信号较完整地保留了原始信号中的信息, 而正常人语音的频率一般在 20 ~ 5 000 Hz 范围内,鉴于此,本文采样频率设置为 16 kHz; 接着,对所得的语音信号进行预处理,包括预加重单元、分帧处理单元、加窗函数单元、端点检测单元; 对预处理后的语音信号进行语音特征的提取; 通过选择,进行 HMM 模型训练或匹配评价; 由匹配输出的概率,转换为分数,作为结果输出.

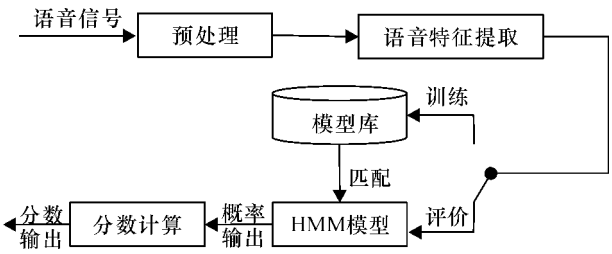

图 1　语音评价过程图

Fig. 1　The process of evaluation

2　语音信号预处理

2.1　预加重

语音信号的平均功率谱受声门激励和口鼻辐射的影响,高频端大约在 800 Hz 以上按 6 dB/oct (倍频程) 衰减,频率越高相应的成分越小,为此要在对语音信号进行分析之前对其高频部分加以提升. 本文采用了 6 dB/oct 的高频提升预加重数字

收稿日期: 2010 - 09 - 10;　**修回日期**: 2010 - 10 - 10
基金项目: 广东省自然科学基金项目(9151042001000017) ; 广东省科技计划项目(2008B080701007) 资助
作者简介: 方凡泉(1956 -) ,男,副译审. E-mail: fqfang@ mail. gdufs. edu. cn
* 通讯作者. E-mail: lin_hao@ foxmail. com

滤波器,实现对语音信号高频部分的提升,使信号的频谱变得平坦,保持在低频到高频的整个频带中,能用同样的信噪求频谱.

2.2 分帧加窗

语音信号具有时变特性,但是在一个短时间范围内,其特性基本保持不变即相对稳定,语音信号的这种特性称为"短时性",这一短段时间一般为10~30 ms. 所以语音信号的分析和处理一般建立在"短时性"的基础上,即进行"短时分析". 对语音信号流采用分帧处理,一般每秒的帧数有30~100 Frames. 具体帧数的确定,视实际情况而定.

分帧既可以采用连续方式,也可采用交叠分帧的方式,由于语音信号之间存在相关性,本文采用半帧交叠分帧的方式. 这样,对于整体的语音信号来讲,分析出的是由每一帧特征参数组成的特征参数时间序列.

2.3 加窗函数

语音信号具有短时平稳性,可以对信号进行分帧处理. 而为实现对语音信号中抽样 n 附近的语音波形加以强调而对波形的其余部分加以减弱,紧接着还要对其加窗处理. 对语音信号的各个短段进行处理,也就是对各个短段进行变换运算. 本文采用应用最为广泛、效果最好的汉明窗(Hamming),对语音信号进行加窗处理.

2.4 端点检测

语音信号处理中的端点检测主要是为了自动检测出语音的起始点及结束点[6]. 本文采用了双门限比较法来进行端点检测. 双门限比较法以短时能量 E 和短时平均过零率 Z 作为特征,结合了 Z 和 E 的优点,使检测更为准确,从而有效地降低了系统的处理时间,提高了系统处理的实时性,并且能排除无声段的噪声干扰,进而提高了语音信号的处理性能.

图2为单词 book 的语音原信号,对其进行预加重后,语音信号为图3情况. 通过对图3中信号与原信号中的频谱特性比较、分析,发现语音信号在预加重后的频谱变得更为平坦,将有效提高语音特征的质量;在图3中,两条较长的竖线分别表示语音信号的起始点与结束点,由此可知,采用双门限法可获得有效语音,避免了其他非语音段或噪声的干扰,从而减少数据处理量,提高了系统的处理性能.

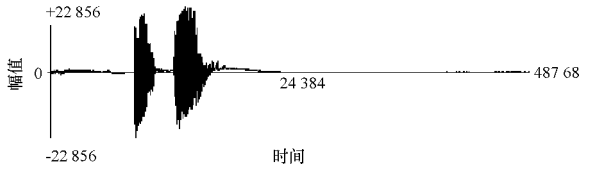

图2 单词 book 语音原信号

Fig 2 Org-signal of word 'book'

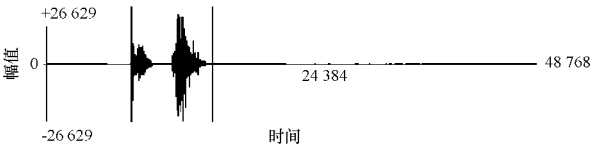

图3 单词 book 语音预加重后及其端点检测情况

Fig 3 After pre-emphasis and endpoint detection

3 语音特征提取

3.1 语音特征

语音信号特征参数提取就是对语音信号进行分析处理,去除对语音评价无关紧要的冗余信息. 原始语音数据因说话人的不同,说话长度、响度等原因,存在太多干扰语义的信息,而且原始语音数据量非常大,不宜直接用于语音质量评价. 所以需要对原始语音数据进行特征提取,最理想的语音特征参数只反映语义信息,而且数据总量小. 特征参数选择的标准应尽量满足以下3点.

(1) 各阶参数之间有良好的独立性;

(2) 能有效代表语音特征,包括声道特征和听觉特征,具有良好的区分性;

(3) 特征参数要计算方便,在保持语音质量的情况下,寻求高效的提取算法方法,以减小存储要求并保证语音评价的实时实现.

3.2 语音特征比较

语音的时域特征主要有语音的短时平均能量和短时平均过零率及基音周期. 所谓基音周期就是声带发声时振动的周期,因为辅音的幅度小,且没有明显的周期性,元音的幅度大,有明显的周期性,所以一般的基音周期指的是语音中元音的周期. 语音的频率特征有多种,常见的有 FFT(Fast Fourier Transformation) 频谱系数、LPC 系数、LPC 倒谱系数(LPCC, Linear Predictive Cepstral Coding)、Mel 倒谱系数(MFCC),等等. 实验证明

LPCC 特征参数与 MFCC 特征参数是较好的表征语音特征的参数. 二者都是将语音从时域变换到倒谱域上,前者利用线性预测编码(LPC) 技术求倒谱系数,后者则直接通过离散傅立叶变换(DFT, Discrete Fourier Transformation) 进行变换.

线性预测分析(LPC) 是较为常用的语音特征分析方法. 由于 LPC 方法有效地解决了短时平稳信号的模型化问题,可把语音信号看成是由全极点产生的,较好的逼近共振峰,提供谱估计,算法简洁准确,计算量小,便于实时处理. 仅用 12/个 LPC 系数就能很好地表示复杂语音信号的特征,大大地降低了信号的冗余度,有效地减少了计算量和存储量,使之成为语音评价的基础. 但是 LPCC 参数假定所处理的信号为 AR 信号,对于动态特性较强的辅音,这个假设并不严格成立,于是将基于声道的 LPCC 特征[7]用于评价过程,存在着一定的局限.

MFCC 参数比 LPCC 参数更符合人耳的听觉特性,具有更高评价的特性[8]. 由于语音的信息大部分集中在低频部分,而高频部分易受环境干扰,MFCC 参数将线性频标转化为 MEL 频标,强调语音的低频信息,从而突出了有利于识别的信息,屏蔽了噪声的干扰. 另外,MFCC 参数无任何前提假设,在各种情况下均可使用,可更客观反映语音特征.

综合以上分析,LPCC 参数是基于线性频标的,而 MFCC 是基于 Mel 频标的,更符合人耳听觉模型,具有更好的客观性,于是本文采用 MFCC 特征作为语音评价的特征.

3.3　MFCC 语音特征提取

Mel 倒谱系数(Mel-Frequency Cepstral Coefficients,MFCC) 是根据人类听觉系统的特性提出的,模拟人耳对不同频率语音的感知. 人耳分辨声音频率的过程就像一种取对数的操作. 例如,在 Mel 频域内,人对音调的感知能力为线性关系,如果两段语音的 Mel 频率差两倍,则人在感知上也差两倍. MFCC 特征提取过程如图 4 所示.

图 4　MFCC 特征提取过程

Fig. 4　Process of MFCC features extraction

图 4 中,首先,对语音信号进行傅里叶变换(FFT) ,将时域转换到频域; 其次,采用 Mel 滤波器组进行滤波; 第三,滤波器输出进行对数操作; 最后经过 DCT(Discrete Cosine Transformation) 运算,输出 MFCC 特征值. 表 1 为单词 book 的 MFCC 特征提取结果.

表 1　单词 "book" 的 MFCC 特征值

Table 1　The MFCC features of word 'book'

Mel 滤波器	MFCC	Mel 滤波器	MFCC
0	9.142 66	7	2.171 26
1	-6.515 78	8	-0.766 243
2	4.070 46	9	-0.300 495
3	-1.684 45	10	1.170 25
4	0.538 736	11	-1.102 05
5	1.333 96	12	0.805 075
6	-2.097 56		

4　评价模型的选择

语音信号经过了预处理、特征提取后,要利用语音特征进行语音质量的评价. 目前,对语音质量评价的模型有动态时间规整(DTW,Dynamic Time Warping) 和隐式马尔可夫模型(HMM, Hidden Markov Model) .

其中,DTW 模型计算量比较小,有利于提高系统的实时性. 但 DTW 模型比较常用于特定人的评价,即只能用于同一人在不同时间对同一段语音内容的评价,不具有广泛性. 另外,男性和女性的声音频带存在着较大的差异,而 DTW 模型不具备进行多人训练的功能,不能综合多人的语音特征进行语音评价,所以采用 DTW 模型进行评价具有很大的局限性.

鉴于此,本文采用具有更好评价性能的 HMM 模型[9-10]. 首先,录制 5 位专家(2 位女性,3 位男性) 的语音,并对每一段语音信号进行预处理; 其次,进行语音特征的提取,对不同专家的语音特征进行标示; 第三,将 5 位专家的语音特征输入 HMM 模型,训练形成具有 5 位专家语音特征的 HMM 模型; 最后,录制待评价的语音,提取其特征后输入 HMM 模型,进行语音特征配对,输出相似度,建立评分机制,将分数相似度转换为分数输出. 图 5 为评价结果图.

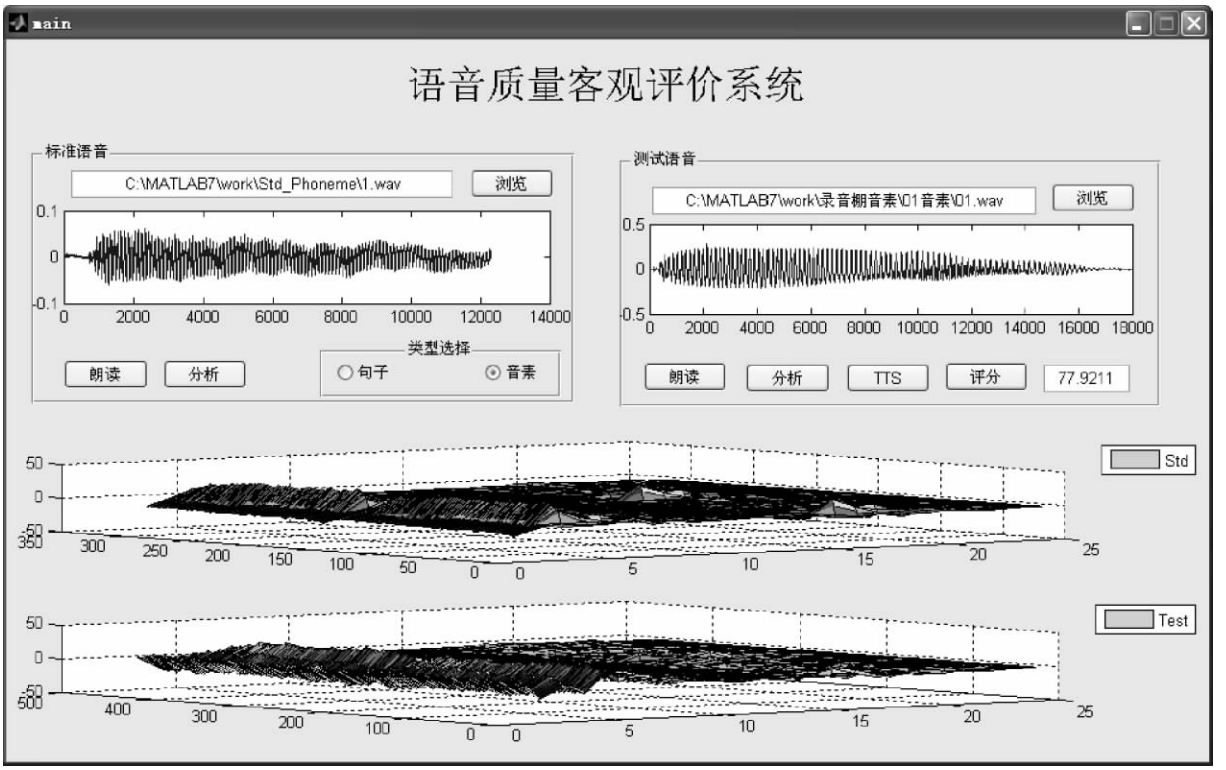

图5 评价结果图

Fig. 5 Interface for evaluation result

图5为音素[iː]的评价结果图. 图中显示了经端点检测后的语音信号,在三维图中,是语音信号的MFCC特征. 通过HMM模型后输出两段语音信号的相似度,通过评分机制,并将结果转换为分数. 表2为音素和单词的评价结果,其中编号代表学生编号,内容即为评价内容,分数采用百分制计算.

表2 音素及单词的评价结果

Table 2 The evaluation results of phonemes and words

编号	内容									
	01	02	03	04	05	06	07	08	09	10
/iː/	77.92	73.42	68.91	76.63	83.52	82.70	82.05	78.94	80.36	81.27
bean	75.35	71.26	65.21	73.56	80.45	79.64	80.57	76.32	78.28	78.98
/uː/	84.92	82.83	78.69	84.81	79.46	73.83	86.42	88.83	85.26	86.42
spoon	82.35	80.62	77.55	85.23	78.32	72.26	83.20	82.21	80.53	85.15
/ɑː/	75.26	74.32	70.58	76.59	80.65	81.36	81.27	79.52	79.53	80.62
barn	74.36	72.53	70.36	73.42	75.62	75.23	80.10	76.29	75.36	78.77
/ɔː/	81.36	80.27	76.82	86.57	85.54	83.57	86.53	76.37	78.56	80.72
born	78.73	76.86	75.58	84.32	83.69	82.91	83.89	74.11	75.23	78.62
/aʊ/	74.56	77.36	70.69	75.78	79.63	77.57	78.82	76.37	75.69	77.82
how	73.29	75.87	71.27	73.59	76.37	76.84	75.73	72.89	74.67	75.79

评价结果表明,同一个人音素与对应单词的评价结果比较接近,且单词分数稍低于音素,主要原因是单词包含了更多语音信息,信息越多,与标准的距离就越大,所以分数也会稍低.

5 结束语

本文针对音素与单词的评价,提出采用MFCC

特征结合 HMM 模型进行评价的方法,并设计系统验证了方法的客观性. 评价结果经学校语音专家验证,具有比较好的评价客观性.

References:

[1] ZHANG Wen-zhong, GUO Jing-jing. Fuzzy marking: A new approach to marking oral proficiency [J]. Modern Foreign Languages, 2002(1): 98-102. (in Chinese)

[2] CHEN G, PARSA V. Objective speech quality evaluation using an adaptive Neuro-Fuzzy Network [J]. Studies in Computational Intelligence, 2008, 97-116.

[3] FRANCO H, NEUMEYER L, YOON K. Automatic pronunciation scoring for language instruction [J]. ICASSP-97, 1997, 1471-1474.

[4] NEUMEYER L, FRANCO H, WEINTTRAUB M. Automatic text-independent pronunciation scoring of foreign language student speech [J]. ICSLP'96, Philadelphia, PA, USA, Oct. 1996, 1457-1460.

[5] HORACIO FRANCO, LEONARDO NEUMERER. Combination of machine scores for automatic grading of pronunciation quality [J]. Speech Communication, 2000, 30: 121-130.

[6] LIU Qing-sheng, XU Xiao-peng, HUANG Wen-hao. Research on a speech endpoint detection method [J]. Computer Engineering, 2003, 29(3): 120-123. (in Chinese)

[7] YU Jian-chao, ZHANG Rui-lin. Speaker recognition method using MFCC and LPCC features [J]. Computer Engineering and Design, 2009, 30(5): 1189-1191. (in Chinese)

[8] WU Zun-jing, CAO Zhi-gang. Improved MFCC-Based feature for robust speaker identification [J]. Tsinghua Science and Technology, 2005: 158-161.

[9] ARNAB G, PAVEL L, SANJEEV K. Hidden Markov models for automatic annotation and content-based retrieval of images and video [J]. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM, 2005: 544-551.

[10] HE Jue, LIU Jia. The optimal selecting for HMM state-number in Mandarin continuous speech [J]. Journal of Chinese Information Processing, 2006, 20(6): 85-90. (in Chinese)

参考文献:

[1] 张文忠,郭晶晶. 模糊评分: 外语口语测试评分新思路 [J]. 现代外语,2002(1): 98-102.

[6] 刘庆升,徐霄鹏,黄文浩. 一种语音端点检测方法的探究 [J]. 计算机工程,2003,29(3): 120-123.

[7] 余建潮,张瑞林. 基于 MFCC 和 LPCC 的说话人识别 [J]. 计算机工程与设计,2009,30(5): 1189-1191.

[10] 何珏,刘加. 汉语连续语音中 HMM 模型状态数优化方法研究 [J]. 中文信息报,2006,20(6): 85-90.

Research and implementation on objective speech quality evaluation

FANG Fan-quan, LI Xin-guang, WANG Gui-zhen, LIN Jiang-hao

(Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: Objective speech quality evaluation plays a pivotal role in SALL(Self-Access Language Learning). This thesis firstly introduces the pre-treatment unit of voice signals in the process of objective speech quality evaluation. Then, it also presents diverse voice features extraction algorithms and makes comparisons between them. After that, the MFCC feature, which is closer to human auditory model, is chosen in the extraction process. Meanwhile, the whole process and its results are provided. Finally, we compare the advantages and disadvantages of current evaluation models(DTW and HMM), and propose an evaluation method which adopts the HMM model, and thus design a system to test and verify the objectivity of evaluation using this method.

Key words: speech quality; objective evaluation; HMM Model; voice features

【责任编辑: 陈　钢】