

基于语音识别技术的英语发音评测系统研究

滕海坤,刘心声,王丽红

(黑河学院 计算机与信息工程学院,黑龙江 黑河 164300)

摘要:利用语音识别技术,依英语发音的特有性,搭建完整合理的计算机辅助英语发音评测系统。整个系统包括 3 个部分:语音段验证、语音信号切割和发音评测。其中语音段验证利用置信度评估的技术,并依据置信度阈值对评测语音内容是否需要评测做出判断;语音信号切割是以预先训练好的英语发音声学模型为切割依据,通过强制对位的方式切割出正确的发音区段;英语发音评测是系统的核心,使用的评测方式是比较评测语音和标准语音的相似度,采用 4 个评测特征:发音完整度、韵律性、语速和准确度进行发音相似度评测。针对不同的特征参数设计合理的评测机制,并对 4 个特征参数进行加权运算,以建立一个完整合理的英语发音评测系统。实验证明,系统能够给出比较客观的评分,达到预期设计目标。

关键词:语音识别;强制对位方式;语音段验证;置信度;评测机制

中图分类号:TP391.42 **文献标识码:**A **文章编号:**1671-5322(2016)01-0017-06

随着云计算的出现和人工智能的广泛应用,在移动互联网浪潮的推动下,智能语音技术取得了突破性发展和创新性应用。在国家倡导的英语教学改革背景下,计算机辅助英语教学 CALL (Computer Aided Language Learning) 应运而生,为学习者提供了更为真实更为便利的英语学习环境。该软件能够有效地指导学习者高效率地学习口语、评价学习者的发音、给予学习者有效的反馈并指导学习者纠正其错误发音,即训练和提高学习者“说”英语的能力^[1]。CALL 系统一般采用语音压缩编码技术、语音识别与合成技术及口语评测技术等。基于语音技术的 CALL 系统研究在国内还处于起步阶段,许多研究者开展了相关研究工作,希望改善英语教学环境^[2-3]。本文主要结合语音识别技术研究计算机辅助英语发音评测系统。

发音评测系统主要包括语言段验证、语言信号切割和发音评测 3 部分。语音段验证作为系统的第一道检视关卡^[4-5],确保了评测语音内容的正确性后;利用隐马尔可夫模型(HMM)的维特比算法(Viterbi algorithm)将评测语音切割出每个元/辅音(以下简称元音)的时间距离或者关联,

使用参数对比的方法(本文使用 4 个对比参数:发音完整性、基音周期(韵律性)、语速和 HMM 对数概率距离(准确度))将评测发音与标准发音语料库(字典)进行逐音对比,计算出评测发音和标准发音之间的差异程度,并根据评分映射模型将差异程度转换为直观的分值。

1 计算机辅助英语发音评测系统架构

本文英语发音评测系统架构流程如图 1 所示。首先对预处理过的学习者英语发音进行语音段验证,包括元音段切割、验证系统的建立和验证系统的可靠性等;然后通过 HMM 模型对大量的标准发音资料库进行声学模型训练,使用 Viterbi 算法对语音段切割解码,再将这些信息包含评比参数的提取、评比参数正则化、参数关联流程和评测机制等送至英语发音评测系统的核心,即发音评测模块,通过该模块求出各评测参数在英语发音评测中的权值,以反映人类专家对英文语句好坏的看法,并向学习者提供的反馈结果,包含综合评分、通过专家知识库对比后的矫正意见^[6]。

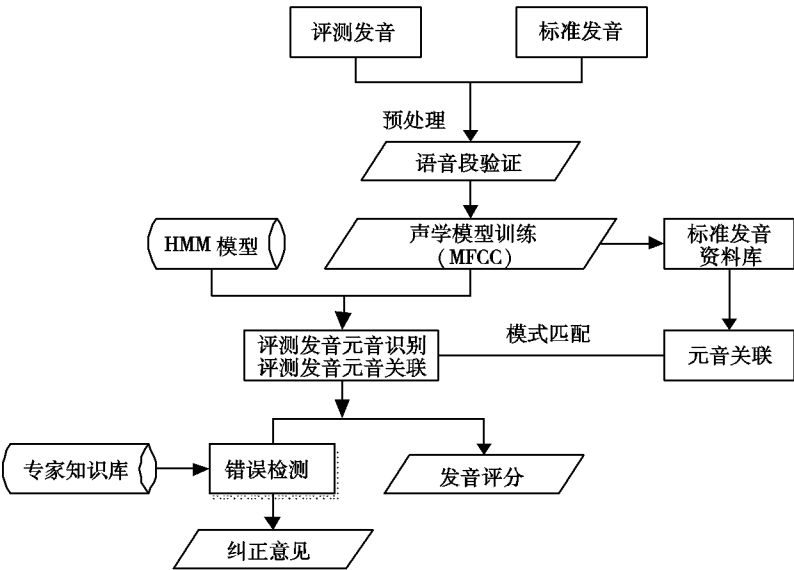


图 1 英语发音评测系统图

Fig. 1 English pronunciation evaluation system diagram

2 语音段验证

所谓的语音段验证 (speech Verification), 就是针对不同的评测语音产生判断阈值, 并依据阈值对该评测语音内容的正确性做出判断。语音段

验证过程如图 2 所示^[7-9], 当验证系统接收到评测语音段后, 分别对每个元音进行模式匹配, 再依匹配结果的距离大小, 结合验证机制给予最后的置信度阈值。

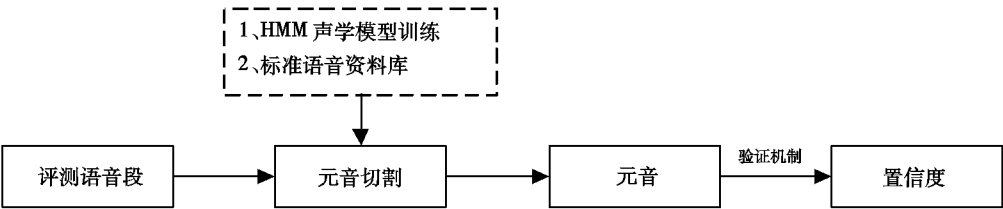


图 2 语音段验证流程

Fig. 2 Speech verification process

2.1 元音切割

本文使用 Viterbi 算法中常用的强制对位 (Forced Alignment) 方式, 将语音段尽可能地切割出最小元音发音段^[10]。在这种情况下, 评测语音切割后, 如果评测内容和标准发音内容相识度较高, 则经切割后产生元音段的数量将接近或等同于标准发音元音段的数量; 如果评测内容前面 n 个元辅音段与标准发音相同, 后面的元辅音段是标准发音库以外的语音段, 则经过强制对位后的元辅音数量大约为 n 个。例如, 如果标准发音为「She had your dark suit in greasy wash water all year」, 评测发音为「She had your dark suit」, 经过

元音段切割后, 系统中能够切割出来的元辅音段个数为 15。对于没有切割出来的部分语音段, 本文将其置信度设置为 0, 以增强语音段验证的可靠性, 使得系统的识别率更高。图 3 为上面两段发音经过切割后产生的不同结果, 其中横轴表示时间 (s), 纵轴表示能量, 蓝色间隔表示一个有效元音段。由图 3 可知, 上半部分发音内容与标准发音相同, 系统能够切割出完整的元音段; 下半部分发音的内容与标准发音有一部分不同, 系统只能切割出标准发音库中的元音段, 不能够切割出库以外的元音段。

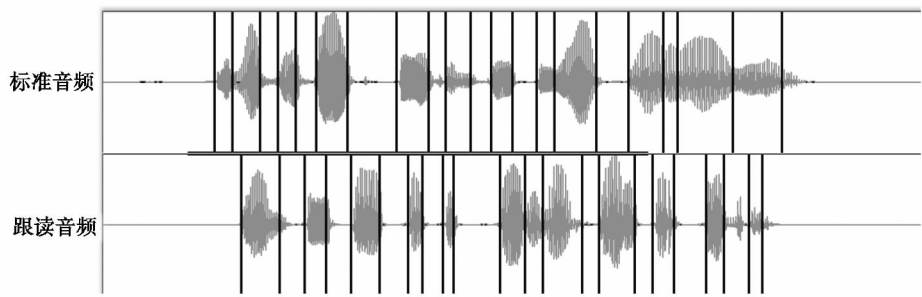


图 3 语音切割对比图

Fig.3 Speech cutting contrast figure

2.2 验证机制

综上所述,语音段经过切割后会出现两种情况:一种是切割后得到的元音段与标准发音的元音段个数基本一样,一种则是部分的语音内容没有切割出元音段。本节主要讨论第一种情况,通过评比参数正则化得到一个适合的置信度阈值。从参考文献[5]中可以知道元音段的排名直接影响到验证系统的置信度值,本文将 SUKKAR 等^[5]提出的求置信度值公式进行改进优化,得到以下公式:

$$C_{\text{Vow}} = \frac{2}{1+e \left(k \cdot R \cdot \frac{(\log_{39}^P R)}{\log_{39}^P R_o} \right)}$$

($\frac{\log_{39}^P R}{\log_{39}^P R_o}$ 表示该元音对数概率值差异度) (1)

C_{vow} 表示置信度值, k 为调整的参数值, $R(0 \sim 38)$ 表示该元音在 39 个模型 (models) 中的排名, 0 表示第 1 名。由公式 (1) 知, 当某元音相对于 39 个模型的排名为第 1 名时, 该元音的置信度值为 1。以 3.1 节例子中【S】元音在语音区段对应于 39 个模型所产生的排名由公式 (1) 换算成置信度值, 如图 4 所示。从图 4 可以看出, 当名次在第 10 名左右时, 置信度值已经降至 0.3 了。

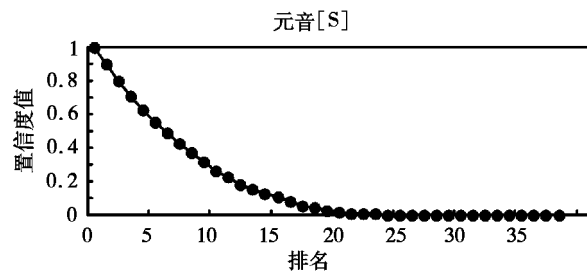


图 4 元音【S】的排名与置信度的关系

Fig.4 Vowels 【S】 ranking relations with confidence

由于元音发音存在相似性, 因此在进行置信度值判断时, 不能只依赖排名进行比较。例如, 【a】和【æ】对比完 39 个模型后名次是第 2 名, 其第 1 名是【ei】。可以看出【a】和第 1 名的对数几率差异度较小, 因此在上述公式中加入了对数概率差异度因子, 保证验证系统的准确性。当计算出语音段中所有元音的置信度值后, 利用每个元音的时间长度 ($T_{\text{len}}(\text{Vowel}_n)$) 和语音段时间长度 ($T_{\text{len}}(\text{Segment})$) 的比值作为权值推导出语音段的置信度值 (0 ~ 100)。公式如下:

$$C_{\text{seg}} = 100 \cdot \sum_{n=1}^N \frac{T_{\text{len}}(\text{Vowel}_n)}{T_{\text{len}}(\text{Segment})} \cdot C_{\text{Vow}_n}$$

N 为单词中元音的个数。

2.3 验证系统实验结果

为验证系统的置信度阈值。本文选取两段语音 (150 句) 进行比对, 一段等同于标准发音内容, 即正确发音段; 一段为不完全等同于标准发音内容, 即错误发音段。语音格式都为 Wav, 取样频率为 16 kHz, 位数为 16 bits。如果语音段得出的置信度值高于阈值, 可以认为评测发音内容和标准发音内容是相同的, 即发音评测结果是有效的; 反之, 系统将停止对两个不相同的内容进行后续的评测打分。两段语音通过验证系统得到相应的置信度值如图 5 所示。

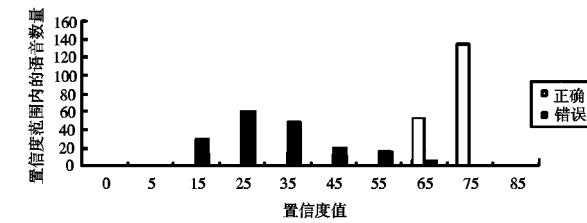


图 5 置信度阈值实验分布情况

Fig.5 Distribution of confidence threshold experiment

由图 5 可以看到正确语音中最小置信度值为 58.64, 错误语音中最大置信度值为 60.00, 通过错误率(ERR) 指标将语音段验证系统的置信度阈值设置为 59.32(两者平均值)。阈值确定后, 对另一组语音段进行验证, 正确语音识别率为 100%, 错误语音识别率为 99.1%。经过算法优化后, 语音段验证系统的各项指标都有所提高。

3 基于声学模型的发音评测

在进行语音切割之前, 要先建立声学模型作为评测参考标准, 继而针对语料中标准发音及评测发音切割出元音段, 然后运用训练好的标准发音模型进行评测打分。

3.1 声学模型设计

声学模型的建立是系统开发工作的数据准备, 模型的好坏取决于语料库, 语料库的好坏直接影响评测结果的准确性。本文依据自录语料库使用 HMM 声学模型对英语标准发音的声学模型和中国人发音的声学模型进行训练, 分别建立标准语料库和评测语料库。中国人发音的声学模型能够代表用户发音特点, 主要作为学习者训练的评测标准, 英语标准发音的声学模型用于发音纠正。在训练语料的特征参数提取中, 本文采用 39 维的 MFCC 特征向量作为声学模型的特征参数, 比较评测发音和标准发音并计算其距离, 以利评测机制的运行^[11]。

目前绝大多数声学模型设计都采用与参考语音模型的相似性为度量的 HMM 模型。常见的 HMM 模型有离散 HMM、连续 HMM 以及离散—连续 HMM 模型。根据实验条件以及客观效果, 本文采用无跨越自左向右的 HMM 模型, 每个元音段模型为 5 个状态, 其模型参数通过 Baum - Welch 算法反复计算得到: $\lambda = f(\pi, A, B)$; 文中 3 个参数初值均设定为等概率初值。由该模型计算某元音段得到的状态转移矩阵如图 6 所示。

$$A = \begin{Bmatrix} 0.979\ 2 & 0.207\ 8 & 0 & 0 & 0 \\ 0 & 0.985\ 4 & 0 & 0 & 0 \\ 0 & 0 & 0.972\ 4 & 0 & 0 \\ 0 & 0 & 0 & 0.969\ 1 & 0.031 \\ 0 & 0 & 0 & 0 & 1 \end{Bmatrix}$$

图 6 状态转移矩阵
Fig.6 State transition matrix

3.2 英语发音评测

英语发音评测是整个学习系统的核心。为了评测结果更加立体化, 本文提出从 4 个方面特征得出评测结果: 发音完整性, 即语音信号是否对短句完整读出; 基音周期, 即语音音高的起伏; 语音急缓变化, 即语音的语速变化; HMM 对数概率, 即语音信号经辨识后得出的每个元音的对数概率, 表示发音内容。最后将各个方面进行综合评分, 每个方面评测权值需要经过理论计算和实验测量确定。

3.2.1 参数对比方式

在上述评测参数提取中, 将切割得到的元音段数量与标准语句数量差异表示完整度, 使用强制对位得到各元音的时间关联(语速变化) 和 HMM 对数概率距离。基音周期的提取一般采用自相关法和基于短时幅度差法, 但效果不理想而且时效性较差。本文根据语音信号时域特征求取基音周期, 其过程为: 1) 求取一帧内信号能量最大点, 并记录其时域位置; 2) 向前或向后搜索次大点位置, 计算和最大点位置距离; 3) 将得出的距离进行标准判断: 首先是正常基音周期范围, 然后继续搜索第 3 大点, 检查距离是否比较接近。经过这种方法求出基音周期数量足够丰富, 能够比较全面地反映发音过程基音周期的变化规律。图 7 为基音求取效果, 横轴表示时间(s), 纵轴表示能量(对应红点表示基音周期值), 红点虚线表示基音周期变化曲线。

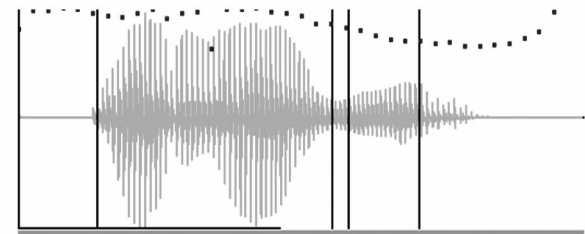


图 7 基音周期求取效果
Fig.7 Pitch period to calculate the effect

在评测参数对比之前, 首先将不同的特征参数提取进行正则化处理, 然后使用 Euclidean Metric(欧几里德距离) 计算特征参数之间的距离, 即差异程度。参数对比流程如图 8 所示。

3.2.2 评测机制建立

对元音 HMM 对数概率、发音完整度、语速及发音韵律等单一特征进行评测, 公式如下: 以此为

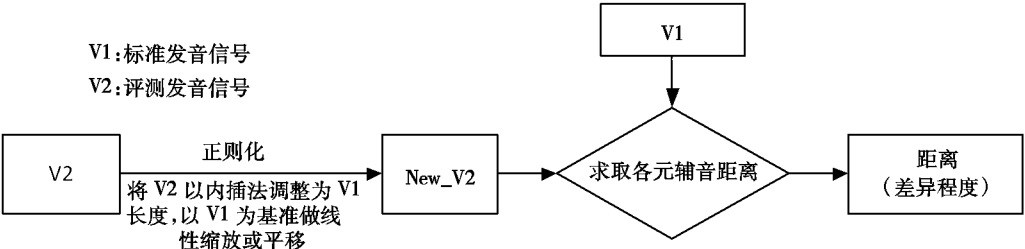


图 8 参数对比流程图

Fig. 8 Parameters contrast flow chart

基础,再对单词和语句进行评测,得到评测的结果。

$$S_{\text{fea}} = 100 \cdot \frac{1}{1 + \alpha \cdot (\text{dist})^\beta} \quad (3)$$

式中:α 和 β 是常量,dist 表示两个元辅音某个特征参数的差异距离,这样可以将 4 个特征参数的差异距离转换成 0 ~ 100 内的分数。只要设置合适的两组参数距离 dist 及对应的 S_{fea} ,即可求出 α 和 β。然后利用 4 个特征参数对评测系统所占的权值计算每个元音的综合特征分值:

$$S_{\text{Vow}} = W_1 \cdot S_{\text{fea}_1} + W_2 \cdot S_{\text{fea}_2} + W_3 \cdot S_{\text{fea}_3} + W_4 \cdot S_{\text{fea}_4} \quad (4)$$

式中 W_1 、 W_2 、 W_3 和 W_4 分别代表 4 个评测参数的权值。在得到每个元音的评分后,以每个元音占单词(Word)的时间比值为权值,求出句子中每个单词的评分。单词评测公式如下:

$$S_{\text{Word}} = \sum_{n=1}^N \frac{T_{\text{len}}(\text{Vowel}_n)}{T_{\text{len}}(\text{Word})} \cdot S_{\text{Vow}_k} \quad (5)$$

式中 N 为一个单词中元辅音的数量。

同理,也可以使用单词在句子中的时间比值为权值计算出句子的评分。

3.2.3 评测实验

通过评测机制得到 4 个评分参数中各元音的差异程度后,依据所占的比例求出一个句子的平均差异程度,即可代入以下公式:

$$S = W_1 \cdot \frac{100}{1 + \alpha_1 \cdot (\text{dist}_1)^{\beta_1}} + W_2 \cdot \frac{100}{1 + \alpha_2 \cdot (\text{dist}_2)^{\beta_2}} + W_3 \cdot \frac{100}{1 + \alpha_3 \cdot (\text{dist}_3)^{\beta_3}} + W_4 \cdot \frac{100}{1 + \alpha_4 \cdot (\text{dist}_4)^{\beta_4}} \quad (6)$$

式中 α_n 和 β_n 是差异程度转化评分的参数, W_n 为 4 个评分参数的权值,dist_n 为标准发音和评测发音讯号在比对后 4 个评分参数的距离。经由

以下实验,即可求得各参数值。

在语料训练部分收集 7 个人 10 个常用短句的语音数据来做实验。采样环境为封闭实验室,采样率为 16 K 单声道。每人每句发音 3 次,其中 4 个人的语音数据用于训练模型,另外 3 个人语音数据用于测试效果。接着请外语老师给每人每一句的发音进行人工打分,并最终给出每个语句的平均得分。通过实验收集了 10 个语料的 4 个特征参数对应的差异程度和综合评分,使用 MATLAB 的 Fminsearch 函数式,利用 Simplex Downhill Search 方法,得到 HMM 对数概率的权值为 60.4%,基音周期的权值为 4.1%,发音完整度的权值为 20.9%,语速的权值为 14.6%。

将 210 个测试语句的人工打分进行分级:高(100 ~ 80)、中(79 ~ 60)、低(59 ~ 0),系统评测结果也按照此方法进行分级。统计人工评测和系统评测的对应关系,结果如表 1 所示(表中数值表示两种评测结果的交叉语句数目)。从表 1 中可以发现,对角线的数值高于每一列的数值,表明本文的评测系统能够客观地反映学习者的发音情况,两者之间的正相关性约为 59%。

表 1 人工评测结果和系统评测结果的对应关系

Table 1 Artificial evaluation results and system evaluation results

系统评测结果	人工评测结果		
	高	中	低
高	65	10	8
中	20	28	25
低	7	16	31

4 结束语

为了使整个系统对结果评测的可信度更具有说服力,本文使用语音段验证对评测语音进行初

步评估,只有评估值高于设定的置信度阈值,评分才具有一定意义。语音信号切割使用强制对位方式得到每个元音的时间关联,使用识别率较高的 HMM 声学模型进行声学模型训练。英语发音评测包括特征参数提取、参数对比的设计和评测机制的建立等。由实验可知,准确度即 HMM 对数概率差异在英语发音评测中占有的重要性最高,而基音周期则最低。英语发音评测的运用相当广

泛且实用,只要搭配良好的设计,学习者可以不断地提高英语的口语能力。在英语学习中,学习者的地方口音和不自然的发音问题总是难以解决。当前,系统中还存在不足,尤其是语句长度对系统评测结果的影响。如何根据语言学习的规律,构建更加有意义的口语学习情境,给出一个专业的反馈意见,将是下一步改进的重点。

参考文献:

- [1] 岳东剑,季洪飞. 语音处理技术在语言学习中的应用[J]. 计算机工程与应用,2001,37(4):112-114.
- [2] 刘庆升. 计算机辅助普通话发音评测关键技术研究[D]. 合肥:中国科学技术大学,2010.
- [3] 赵博,檀晓红. 基于语音识别技术的英语口语教学系统[J]. 计算机应用,2009,29(3):761-763.
- [4] ARIAS J P, YOMA N B, VIVANCO H. Automatic intonation assessment for computer aided language learning[J]. Speech Communication,2010,52(3):254-267.
- [5] SUKKAR R A, LEE C H. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Sub-word based Speech Recognition[J]. IEEE Transactions on Speech and Audio Processing,1996,4(6):420-429.
- [6] 李艳玲,颜永红. 多特征融合的英语口语考试自动评分系统的研究[J]. 电子与信息学报,2012,34(9):2 097-2 102.
- [7] 涂惠燕,陈一宁. 基于语音识别和手机平台的英语口语发音学习系统[J]. 计算机应用与软件,2011,28(9):64-66.
- [8] 宋芳芳,宋晓丽,马青玉. 基于语音识别技术的英语口语自学系统评分机制的研究[J]. 电脑知识与技术,2009,5(7):1 726-1 728.
- [9] 李尘一. 基于联合得分的语音置信度评估系统的研究与设计[D]. 呼和浩特:内蒙古大学,2006.
- [10] 黄骁勇,虞维平. 语音识别技术在外语口语学习中的应用[J]. 计算机系统应用,2006(6):18-21.
- [11] BUTCHER P G, JORDAN S E. A comparison of human and computer marking of short free-text student responses[J]. Computers & Education,2010,55(2):489-499.

The Research of English Pronunciation Evaluation System Based on the Speech Recognition Technology

TENG Haikun, LIU Xinsheng, WANG Lihong

(Heihe University, Computer Science and Information Engineering, Heihe Heilongjiang 164300, China)

Abstract: using the speech recognition technology, according to the characteristic of English pronunciation, build a complete, reasonable evaluation system of computer aided English pronunciation. The whole system includes three parts: speech signal cutting, speech verification and pronunciation evaluation. The speech signal cutting based on prior training good English pronunciation as cutting based on acoustic model, through the way of mandatory registration section cut out the correct pronunciation. Speech segment validation using technology of confidence level evaluation, and on the basis of confidence threshold to judge evaluation speech content evaluation. English pronunciation evaluation is the core of the system, using the way of evaluation was to compare the similarity evaluation speech and standard speech. This paper uses four evaluation characteristics: pronunciation integrity, rhythm, speed and accuracy of the pronunciation similarity measure. According to the characteristics of the different parameters design and reasonable evaluation mechanism, and the four characteristic parameters are weighted arithmetic, to establish a complete and reasonable evaluation system of English pronunciation. Via experiment proves that the system can give more objective grading, achieved the desired design goal.

Keywords: speech recognition; Mandatory registration; Speech segment validation; Confidence; Evaluation mechanism

(责任编辑:李华云)