ISSN 1000-0054
CN 11-2223/N
清华大学学报（自然科学版）2005 年 第 45 卷 第 1 期
J Tsinghua Univ (Sci & Tech), 2005, Vol. 45, No. 1

基于音素的发音质量评价算法

梁维谦， 王国梁， 刘 加， 刘润生

（清华大学 电子工程系，北京 100084）

摘 要：面对广大的外语学习者，计算机辅助语言学习系统已经成为一种最佳的口语学习方式。该文提出了一种新的应用于计算机辅助语言学习系统的面向英语学习人群的发音质量评价算法，名为 PASS(phone-based automatic score for l2 speech quality)。PASS 算法以基于隐含 Markov 模型的语音识别和口音自适应技术为基础，考察了音素发音的准确性和流利性信息，定义了音素级的发音质量分数，从而可以综合得到整句的评分结果。在实验室自行采集和精细标注的非母语语音库上与其他评分算法进行比较实验，PASS 与专家评分的句子级相关性达到了 0.66，优于其他算法。目前 PASS 算法已经被成功地应用于清华大学出版社的互动式语言学习系统中。

关键词：语音信号处理；发音质量评价；隐含 Markov 模型；置信测度；语音识别

中图分类号：TN 912.3 **文献标识码**：A

文章编号：1000-0054(2005)01-0005-04

Phone-based pronunciation quality assessment algorithm

LIANG Weiqian, WANG Guoliang, LIU Jia, LIU Runsheng

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: This paper presents an algorithm which can provide pronunciation quality assessments to foreign language learners in Computer-Assisted Language Learning (CALL) systems. The approach is based on non-native speech recognition and accent adaptation techniques using hidden Markov models. Confidence measures at the phoneme level are calculated to check the phone accuracy and speech fluency. The sentence pronunciation score is obtained by analyzing the phone quality results. The algorithm was evaluated using a corpus of non-native speech. Test results using this database show that the approach outperforms other assessment algorithms on correlations with expert scores at the sentence level. The system has been successfully adopted by the Tsinghua University Press interactive language learning system.

Key words: speech signal processing; pronunciation quality assessment; hidden Markov models; confidence measure; speech recognition

面对广泛而又迫切的口语教学需求，目前的师资和传统教学手段都难以满足。解决这一问题的最佳方案是利用计算机辅助外语教学。文中的口语发音质量评价算法正是计算机辅助语言学习系统中的核心技术，其主要基于语音识别技术。

一般对发音质量的评价主要是考察学习者发音的语段特征和超语段特征[1]。本文的算法主要是针对学习者发音的语段特征进行自动评价。评分方法基于音素，与学习内容无关。目前这种文本无关的评价算法一般都是利用标准发音的统计模型结合置信测度算法实现的[1-3]。统计模型主要采用语音识别中经典的隐含马尔可夫模型(hidden Markov model, HMM)。置信测度算法主要是综合语音识别器输出的识别结果来得到最终的发音质量分数。在得到机器评分后，通常是通过把它与专家组的评分结果做相关性测试来评价它的性能。

本文提出了一种新的基于音素的发音质量评价算法 PASS(phone-based automatic score for l2 speech quality)。PASS 在音素级综合语音识别器输出的对数似然比和语速归一化段长信息，并采用马氏距离定义了一种新的发音质量分数。在对学习者发音进行整体评价时，作者对音素分数做了期望和标准差的统计分析从而得到整句发音的评分结果。本文还采用了发音质量主观评价分数较高的非母语语音对 HMM 模型进行了口音自适应。在实验室自行采集的非母语测试语音库上，PASS 算法取得了优于其他评分算法的效果。

1 语音数据库

本文针对的是汉语普通话人群学习英语发音的

收稿日期：2004-06-03
基金项目：国家自然科学基金资助项目（60272016）
作者简介：梁维谦(1977-)，男(汉)，黑龙江，博士研究生。
通讯联系人：刘加，教授，E-mail：liuj@tsinghua.edu.cn

自动评分问题,为此我们需要用于训练的母语语音库,以及用于自适应和测试的非母语语音库。本文选用了英语语音识别中常用的 TIMIT 语音库作为母语训练库。实验室还自行采集了一个非母语语音库,简记为 ME。ME 分为两个部分:一部分包含连续语音,有 116 句不同的文本,共有 34 个说话人,平均每人读 30 句,简记为 ME01;另一部分包含孤立词语音,有 666 个不同的单词,35 个说话人,每人读一遍所有单词,简记为 ME02。语音采集是在安静实验室条件下进行的。为模拟真实应用,要求学习者以参加口语考试的方式进行录音。

作者邀请了清华大学外语系的 3 位老师(简记为 H1、H2 和 H3)对两次采集过程进行了指导,并对语音进行了句子级的评分标注。专家的句子级评分之间的相关性如表 1 所示。表 2 给出了专家评分区间的分布情况。据此,作者分别在 ME01 和 ME02 数据库中挑选出发音分数最高的 5 男 5 女,他们的语音用作 HMM 模型的非母语口音自适应,这部分语音简记为 MEG。并且在剔除 MEG 的前提下,把 ME01 中发音质量在 3 分以上的语音作为非母语语音识别的测试集,共 403 句语音,记为 ME01_T。

表 1 发音质量的专家评分之间的相关性

库名	H1-H2	H1-H3	H2-H3	平均相关系数
ME01	0.79	0.79	0.79	0.79
ME02	0.70	0.66	0.70	0.69

通常学习者的语音越长,能够反映其发音质量的信息也就越多,这样得到的评分结果(无论是专家评分还是自动评分)也就越稳健。这一点符合人类的认知规律。此外,从实验数据上可以看出,机器评分与专家评分的相关系数比专家之间的相关系数低了约 17%。

表 2 发音质量的专家评分区间的分布

分数区间	1	(1,2]	(2,3]	(3,4]	(4,5]
分布	3.8	15.8	21.3	43.3	15.8

2 PASS 算法

PASS 算法的系统框图如图 1 所示。首先对学习者的语音提取声学特征,本文选用了 MFCC 和对数能量特征,以及它们的一阶、二阶差分特征,共 39 维。其后,根据学习者需要学习的内容将音素 HMM 模型拼接成一个强制线性匹配网络,简记为 FA。同时还生成了一个无语法模型限制的音素循环识别网络,简记为 PL。FA 网络的输出表征学习者真实发音与应发标准语音模型的匹配情况。PL 网络是构造置信测度所需的背景模型,其输出表征学习者真实发音的音素级识别结果。最后利用一个发音质量评价模块分析 FA 和 PL 网络输出的音素分割、似然分数和段长信息,计算发音质量分数。如引言所述,本文工作主要在于对非母语 HMM 模型和基于置信测度的发音质量评分算法的研究。下面我们将分别对这两个方面做详细描述。

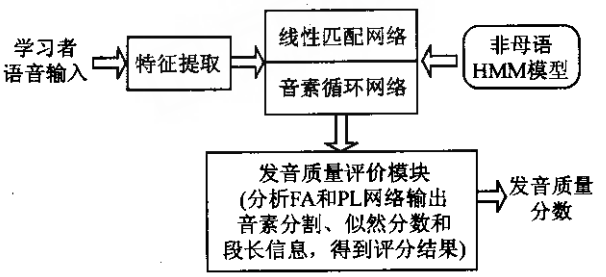

图 1 PASS 算法的系统框图

2.1 非母语 HMM 模型

非母语问题是近年来语音识别领域研究的一个热点。通常的语音识别系统在遇到非母语语音时识别率会明显下降,其原因在于由母语语音训练得到的声学模型与非母语语音之间存在着严重的失配问题。目前解决这一问题的方法主要有两类:一是直接用非母语语音去训练声学模型,二是对现有的声学模型进行非母语自适应。自适应可以是多方面的,如音素集合的自适应,发音习惯的自适应,声学模型的自适应等等[4]。本文采用了 MLLR(maximum likelihood linear regression)与 MAP(maximum a posteriori)级联的自适应方法。

作者采用了英国剑桥大学的 HTK[5]工具包作为语音识别和非母语自适应算法的研究平台。首先将 TIMIT 的 61 音素映射成 CMU 字典[6]的 41 音素,分别训练单音子和状态捆绑的三音子模型。最终单音子模型 41 个,共 120 个状态,每个状态有 16 个对角协方差矩阵的高斯分量。真实存在的三音子模型有 1 185 个,共 491 个状态,每个状态有 8 个对角协方差矩阵的高斯分量。本文用三音子模型作为 FA 网络的声学模型,单音子模型作为 PL 网络的声学模型。

HMM 模型的非母语自适应过程如下:对训练得到的母语 HMM 模型,首先利用 MEG 语音对 HMM 模型的均值参数进行 MLLR 的全局自适应。这里采用全局自适应主要是基于两个方面的考虑:一是目前用作口音自适应的语音数据还不充足,做分类的

MLLR 不够稳健；二是由于非母语与母语发音之间的差异，按照母语模型得到的分类结果并不适用于非母语，从而也会影响分类 MLLR 的稳健性。然后，对得到的 HMM 模型再利用 MEG 语音进行 MAP 自适应。这样做主要是希望能够充分利用有限的非母语语音，进一步消除口音之间的系统差异。

2.2 构造发音质量分数

本文通过 FA 和 PL 网络得到如图 2 所示的音素分割信息。图 2 中，从上到下依次为语音波形、语音的句子级标注、单词级分割信息、FA 网络得到的音素分割和 PL 网络得到的音素识别结果。在每一个音素段，计算了段长归一化对数似然比分数和语速归一化段长分数。

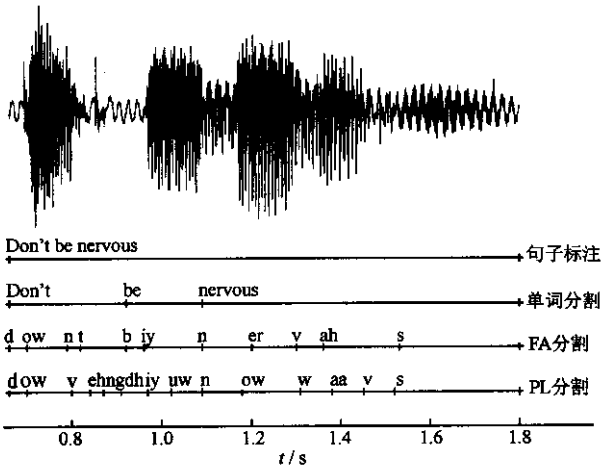

图 2 语音分割、似然比分数与段长分数的示例

1）段长归一化对数似然比

$$R_i = \frac{|L_i^{\mathrm{FA}} - L_i^{\mathrm{PL}}|}{d_i}, \tag{1}$$

其中：下标 i 表示学习者须读语音的第 i 个音素，L_i^{FA} 和 L_i^{PL} 分别为第 i 音素的强制线性匹配对数似然和在第 i 音素的时间段内音素识别得到的对数似然，d_i 为第 i 音素的段长。即 R_i 是以 FA 网络的音素分割为准，得到的 FA 与 PL 似然分数比的段长归一化结果。

2）语速归一化段长

$$D_i = d_i \frac{N}{\sum_{i=1}^{N} d_i}, \tag{2}$$

其中：N 为应读句子的音素数目，D_i 为各个音素发音的相对语速，其表征了发音的流利性。

为进一步提高发音质量评分的可靠性，本文在得到以上两个发音特征的基础上，引入了马氏距离来表示非母语发音与母语发音之间的差距：

$$\lambda(x) = -\frac{|x - (\mu + b)|}{a\sigma}. \tag{3}$$

其中：x 代表 R 或 D 发音特征；μ 和 σ 分别表示 x 的统计期望和标准差；a 和 b 分别为 σ 和 μ 的矫正因子，需实验确定。将图 1 中的非母语 HMM 模型换为母语 HMM 模型，将学习者的语音改为标准的 TIMIT 语音，就能够得到母语说话人的 R 和 D 发音特征。由此可见，$\lambda(x)$代表了真实发音 x 与标准发音之间的距离；$\lambda(x)$越大，表示学习者的语音越接近标准发音，发音质量越好。

此外，为使 $\lambda(x)$与专家评分具有可比性，利用 Sigma 函数，如式(4)所示，将 $\lambda(x)$的值域映射到[1,5]区间。

$$m(\lambda(x)) = \mathrm{Max}\left[\frac{5}{1 + \exp(-\alpha\lambda(x) + \beta)}, 1\right], \tag{4}$$

式中 α 和 β 为经验参数。

进一步将 R 和 D 分数进行线性组合，得到了每个音素的发音质量分数

$$s_i = \frac{1}{2}[m(\lambda(R_i)) + m(\lambda(D_i))]. \tag{5}$$

在计算整句的发音质量分数时，首先计算 $\lambda(R_i)$的均值和标准差，然后相除得到 R_{S}，如式(6)所示。同样对 $\lambda(D_i)$得到 D_{S}。将 R_{S} 和 D_{S} 分别利用式(4)进行映射，再利用式(5)相加，得到句子发音的评分结果。观察式(6)，可以发现，R_{S} 与 $\lambda(R_i)$的均值成正比，与标准差成反比。这说明只有当学习者的每个音素发音质量的平均水平高且波动小时，才认为整句的发音质量好。

$$R_{\mathrm{S}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\lambda(R_i)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\lambda(R_i) - \frac{1}{N}\sum_{i=1}^{N}\lambda(R_i)\right)^2}}. \tag{6}$$

3 PASS 算法的实验与讨论

采用第 1 节所述的非母语语音库 ME01 和 ME02，首先将 PASS 算法与专家评分进行了相关性的比较实验，其次将 PASS 算法与现有的其他一些自动评分算法进行了比较，最后考察了非母语 HMM 模型的识别与评分性能。

表 3 的实验结果表明 PASS 算法在 ME01 上与专家评分的相关性要高于 ME02。这与表 1 中专家评分的相关性结果相一致。

表 3 PASS 评分与专家评分的相关性比较

H1	H2	H3	语音库	平均相关系数
0.64	0.67	0.67	ME01	0.66
0.55	0.54	0.59	ME02	0.56

在表 4 中,将 PASS 算法与文[1]的 GOP 分数(记为 G1)和文[2]的对数似然分数(记为 G2)、后验概率分数(记为 G3)、归一化段长概率分数(记为 G4)和语速分数(记为 G5),在 ME01 上进行了比较实验。实验结果表明 PASS 算法略优于 G1 和 G3,且明显优于其他的评分方法。这主要是因为 PASS 算法综合了似然比和段长两方面的信息,并且引入了母语发音的统计信息作为评分的参照。此外,观察实验数据,可以看到似然比分数对发音质量评分的作用要高于其他参数。

表 4 PASS 算法与其他自动评分算法在 ME01 上的比较

算法	PASS	G1	G2	G3	G4	G5
相关系数	0.66	0.65	0.46	0.64	0.35	0.41

最后在表 5 中对母语和非母语 HMM 单音子和三音子模型进行了性能测试。表 5 中,第 2、3 列分别为音素识别器在 TIMIT 语音库的测试集和 ME01_T 测试集上的识别率,第 4 列为 PASS 评分与专家评分的相关系数。实验结果表明无论是在非母语语音识别还是在发音质量评分上,非母语 HMM 都明显优于传统的 HMM 模型。虽然如此,我们也看到非母语语音识别的识别率与母语识别相比还存在着比较大的差距。本文对非母语 HMM 模型的研究还是比较简单的,这也在一定程度上影响了 PASS 算法的评分效果。

表 5 母语与非母语 HMM 模型的性能比较

HMM 模型	TIMIT 测试集	ME01_T	ME01
母语单音子	62.6	40.7	0.56
母语三音子	68.9	38.7	
非母语单音子	—	48.0	0.65
非母语三音子	—	56.9	

4 结 论

本文提出了一种新的非母语发音质量评分算法,简记为 PASS。PASS 在语音识别阶段采用非母语 HMM 模型对语音进行 Viterbi 分割;在置信测度阶段采用段长归一化对数似然比和语速归一化段长两个发音特征,通过马氏距离定义了真实发音与标准发音之间的距离。从而得到音素和整句的发音质量分数。与其他的自动评分方法相比,PASS 与专家评分之间的相关性更高。PASS 算法可以广泛地应用于计算机辅助口语学习系统和口语测试系统中,使外语学习变得更加方便快捷。

目前 PASS 算法还不是十分令人满意。还存在许多问题值得进一步研究,例如:合理定义发音质量、音素分割的精确性、非母语口音的检测与自适应、发音的韵律特征等等。所有这一切都是今后工作努力的方向。

参考文献 (References)

[1] Witt S M. Use of Speech Recognition in Computer-Assisted Language Learning [D]. Cambridge: The University of Cambridge, 1999.

[2] Franco H, Neumeyer L, Digalakis V, et al. Combination of machine scores for automatic grading of pronunciation quality [J]. *Speech Communication*, 2000, (2-3): 121-130.

[3] Kawai G, Hirose K. A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training [A]. Proceedings of ICSLP [C]. Sydney: IEEE, 1998. 1823-1826.

[4] Tomokiyo M L. Recognizing Nonnative Speech: Characterizing and Adapting to Non-native Usage in Speech Recognition [D]. Pittsburgh: Carnegie Mellon University, 2001.

[5] Young S, Evermann G, Kershaw D, et al. The HTK Book (for HTK Version 3.2) [EB/OL]. http://htk.eng.cam.ac.uk/, 2002.

[6] Weide R L. The CMU Pronouncing Dictionary [EB/OL]. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1998.

基于音素的发音质量评价算法

作者： 梁维谦， 王国梁， 刘加， 刘润生， LIANG Weiqian， Wang Guoliang， LIU Jia， Liu Runsheng

作者单位： 清华大学,电子工程系,北京,100084

刊名： 清华大学学报（自然科学版） ISTIC EI PKU

英文刊名： JOURNAL OF TSINGHUA UNIVERSITY（SCIENCE AND TECHNOLOGY）

年，卷(期)： 2005，45(1)

被引用次数： 8次

参考文献(6条)

1. Witt S M Use of Speech Recognition in Computer-Assiste Language Learning 1999

2. Franco H. Neumeyer L. Digalakis V Combination of machine scores for automatic grading of pronunciation quality 2000(23)

3. Kawai G. Hirose K A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training 1998

4. Tomokiyo M L Recognizing Nonnative Speech: Characterizing an Aapting to Non-native Usage in Speech Recognition 2001

5. Young S. Evermann G. Kershaw D The HTK Book (for HTK Version 3.2) 2002

6. Weide R L The CMU Pronouncing Dictionary 1998

相似文献(1条)

1. 学位论文 李婧 基于UBM的发音质量评价系统的设计与实现 2007

本文将语音信号处理技术应用在外语语言学习中，设计实现一个可以自动评价中国人英语发音质量的系统。该系统集中了人类发音专家的知识，可以自动比较学习者的发音与标准发音之间的差别，并将比较结果以分制或等级的形式反馈给学习者。

本文将全局背景模型(UBM)引入到发音质量评价算法中，提出了一种新的评价发音质量的测度，称为基于UBM的对数似然比。本文采用英国剑桥大学的HTK工具包作为实验平台，验证各种评分算法的有效性。实验证明，在实验室自行采集的干净的非母语语音测试集上，基于UBM的对数似然比分数与专家评分的相关性明显优于其它的评分测度。

为了将中国人英语发音质量评价系统投入实用，本文利用VC 6.0实现了一个完整的发音质量评价系统。该系统实现了三种发音质量测度，分别是对数似然度分数(局部规整)、语速归一化段长分数和基于UBM的对数似然比分数；并利用实验求得的线性回归方程将三者综合起来，得到最终的发音质量评价分数。

引证文献(7条)

1. 颜永红 语言声学的最新应用[期刊论文]-声学学报 2010(2)

2. 颜永红 语言声学进展及其应用[期刊论文]-应用声学 2009(2)

3. 刘庆升. 魏思. 胡郁. 王仁华 基于KLD差的统计错误模式生成算法[期刊论文]-数据采集与处理 2009(1)

4. 李婧. 黄双. 张波 基于UBM的发音质量评价算法[期刊论文]-计算机工程 2008(22)

5. 刘庆升. 魏思. 胡郁. 郭武. 王仁华 基于语言学知识的发音质量评价算法改进[期刊论文]-中文信息学报 2007(4)

6. 黄骁勇. 虞维平 语音识别技术在外语口语学习中的应用[期刊论文]-计算机系统应用 2006(6)

7. 董滨 计算机辅助汉语普通话学习和客观测试方法的研究[学位论文]博士 2006