第34卷 第22期 计 算 机 工 程 2008年11月
Vol.34 No.22 Computer Engineering November 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)22—0207—03 文献标识码：A 中图分类号：TP312

基于UBM的发音质量评价算法

李 婧[1]，黄 双[2]，张 波[2]
(1. 天津理工大学计算机科学与技术学院，天津 300191；2. 南开大学软件学院，天津 300071)

摘 要：将已经成功应用到说话人识别/确认领域中的高斯混合模型和全局背景模型(UBM)引入语音发音质量评价领域，提出一种新的评价英语发音质量的算法。该算法训练出标准发音的全局背景模型。UBM模型描述与音素无关的特征分布，定义段时长归一化的相似度比例对数为音素的发音质量分数，综合得到整句发音的评分结果。实验证明，在实验室自行采集的非母语语音数据库上，该算法评分与专家评分的相关性达到了0.700，优于其他评分算法。

关键词：全局背景模型；对数似然比；高斯混合模型；发音质量评价

Pronunciation Quality Scoring Algorithm Based on Universal Background Model

LI Jing[1], HUANG Shuang[2], ZHANG Bo[2]
(1. School of Computer Science and Technology, Tianjin University of Technology, Tianjin 300191;
2. College of Software, Nankai University, Tianjin 300071)

【Abstract】 This paper presents a new algorithm which can assess the pronunciation quality of the English spoken by Chinese students. The new algorithm uses Gaussian Mixture Model(GMM) and Universal Background Model(UBM), which is successfully used in speaker verification. It calculates the duration normalized log-likelihood ratio of each phone as phonemic pronunciation scores. It combines each phonemic score to obtain the overall pronunciation quality. The algorithm is evaluated by using a corpus of non-native speech. Experimental results show that the approach outperforms other assessment algorithms on correlations with expert scores at the sentence level. In the test database, this method obtains high correlation(0.700).

【Key words】 Universal Background Model(UBM); log-likelihood ratio; Gaussian Mixture Model(GMM); pronunciation quality scoring

1 概述

在语言学习和教学中，反馈占有独特的地位。随着语音信号处理技术的迅猛发展，计算机可以在语言学习中实时提供有效的反馈。目前，已经有一些结合语音识别引擎来训练和提高学习者口语能力的语言学习系统，例如微软研发的Encarta Interactive English Learning、Hebron Soft公司的Caroline in the City/CNN Interactive English，这些系统计算学习者发音和标准发音间的相似度，为用户提供分值或等级等简洁直观的反馈。在语音发音质量评价上，国内外很多研究机构提出了各种评价算法，常见的有对数似然度评分算法[1]、后验概率评分算法[1-2]、段时长评分算法[1-2]。改进的算法有剑桥大学提出的GOP算法[3]、清华大学提出的PASS算法[4]等。

高斯混合模型(Gaussian Mixture Model, GMM)和全局背景模型(Universal Background Model, UBM)现已成功应用于说话人确认/识别系统中，训练用来表示说话人无关的特征分布。本文将GMM-UBM引入语音发音质量评价中，提出了一种新的评价发音质量的算法。实验证明，在实验室自行采集的非母语测试集上，该算法优于其他评分算法。

2 算法原理

2.1 对数似然比

在说话人确认系统中，对于给定的一段语音Y和说话人S，通过如下的对数似然比确认Y是否是S的发音[5]：

$$R=\frac{p(Y|H_0)}{p(Y|H_1)}\begin{cases}\geq\theta & S\\ <\theta & \overline{S}\end{cases} \tag{1}$$

其中，H_0是说话人S的发音数据训练得到的模型；H_1是除S以外的说话人的发音数据训练得到的模型；对数似然比R决定了语音Y和说话人S的相似度，R越大，Y越接近S，反之，Y是S发音的概率越小。同理，对于给定的一段语音Y和一个音素q，应用式(1)求出对应的对数似然比，值越大，说明发音Y越准确，值越小，发音越不准确。

2.2 GMM-UBM模型

在说话人识别/确认系统中，模型H_0用说话人S的GMM来表征，H_1则用UBM表示，它是由数据库中所有说话人的语音训练得到的。将GMM-UBM引入语音发音质量评价中，为了更精确地描述音素q的发音，用隐马尔科夫模型(HMM)表征音素q的发音模型H_0，H_1仍然用由所有音素的发音数据训练而成的UBM表示。在介绍发音质量评价前，先介绍GMM和UBM。

GMM是一种多维概率密度函数，是一个具有M个高斯混合成分的D维高斯混合模型，可以用M个高斯成员的加权

作者简介：李 婧(1979－)，女，助教、硕士，主研方向：语音信号处理；黄 双，硕士；张 波，副教授、博士
收稿日期：2007-12-17 **E-mail：**li_joey@126.com

和表示，即

$$P(\boldsymbol{x}_t \mid \lambda)=\sum_{i=1}^{M} w_i\, p[\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i] \tag{2}$$

其中， $\boldsymbol{x}_t$ 是一个 D 维的观察矢量； $w_i(i=1,2,\cdots,M)$ 为混合权值，相当于每个高斯成员出现的概率，且 $\sum_{i=1}^{M} w_i=1$ ；$p[\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i]$ 为第 i 个高斯函数，定义如下：

$$p[\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i]=\frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)\right\} \tag{3}$$

其中， $\boldsymbol{\mu}_i$ 为均值矢量， $\boldsymbol{\Sigma}_i$ 为协方差矩阵。共有 M 个高斯分布函数，其参数为 $\boldsymbol{\mu}_i$， $\boldsymbol{\Sigma}_i(i=1,2,\cdots,M)$ 。每个函数受到 w_i 加权后，取和得到 $\boldsymbol{x}_t$ 的概率分布。

整个高斯混合模型可以由各均值矢量、协方差矩阵及混合分量的权值来描述，因此，得到一个用三元组表达的模型参数 λ ：

$$\lambda=\{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \quad i=1,2,\cdots,M \tag{4}$$

为了节省计算量，协方差矩阵 $\boldsymbol{\Sigma}_i$ 在实际应用中一般简化为对角矩阵。

UBM本质上是一个大型的GMM[6]。在说话人识别领域，用所有说话人的发音数据训练得到，表征与说话人无关的特征分布；在音素发音质量评价领域，它用所有音素的发音数据训练而成，表征与音素无关的特征分布。

2.3 音素发音评价分数

根据式(1)，对于给定的音素发音 $\boldsymbol{x}_t$ ，定义分数 R_i 来表征音素 q 的发音质量：

$$R_i=\frac{\mathrm{lb}P(\boldsymbol{x}_t \mid \lambda_{phn})-\mathrm{lb}P(\boldsymbol{x}_t \mid \lambda_{ubm})}{d_i} \tag{5}$$

其中，λ_{phn} 为音素q的标准发音HMM模型；λ_{ubm} 为全局背景模型。为了消除音素时长对 R_i 的影响，对 R_i 进行段时长归一化操作，将 R_i 除以音素的发音时长d_p，得到段时长归一化后音素的发音质量评价分数。

在得到语句中每个音素的 R_i 之后，可按照一定的加权规则求取整句的发音质量分数。常用的加权方法有如下 2 种[7]：

$$S_1=\frac{1}{L}\sum_{i=1}^{N} d_i R_i \tag{6}$$

$$S_2=\frac{1}{N}\sum_{i=1}^{N} R_i \tag{7}$$

其中， S_1 将所有音素的发音质量分数进行段时长加权平均； L 是语句中所有非静音语音的帧总数，即 $L=\sum_{i=1}^{N} d_i$ ； S_2 则不考虑每个音素所占的时间长度，直接对所有音素的 R_i 求取平均值。

2.4 算法流程

图 1 描述了算法的基本流程：(1)对学习者输入的语音进行前端处理，包括预处理和特征提取。实验选用了 MFCC 和对数能量特征以及它们的一阶、二阶差分特征，共 39 维。(2)将这段语音输入到强制对齐网络中进行 Viterbi 解码，该网络由学习者的发音脚本组建而成。学习者语音通过 Viterbi 解码后，可得到每个音素的发音起始点、终止点，即得到每个音素对应的发音段 $\boldsymbol{x}_t$ ，此外，还可以得到 $\boldsymbol{x}_t$ 在音素的标准发音模型 λ_{phn} 下的输出概率，即 $\mathrm{lb}P(\boldsymbol{x}_t \mid \lambda_{phn})$ 。(3)将每个音素对应的发音段 $\boldsymbol{x}_t$ 输入到 UBM 模型中得到 $\mathrm{lb}P(\boldsymbol{x}_t \mid \lambda_{ubm})$ 。(4)根据式(5)~式(7)得到整句的发音质量分数。

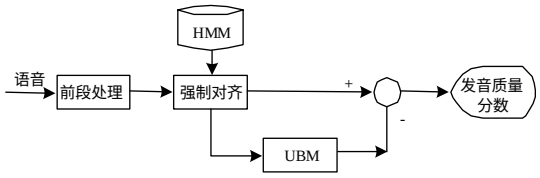

图1 算法原理

3 实验设计

3.1 实验数据库

为了实现自动评测中国学生的英语发音质量，在实验中，需要训练母语发音模型的母语标准发音数据库、自适应数据库和非母语测试数据库。标准发音数据库采用英语语音识别中常用的 TIMIT 数据库，自适应数据库和非母语测试数据库由实验室采集整理而成。其中，自适应数据库中包括 10 个说话人(5 男 5 女)，平均每人读 200 个单词和 30 个句子；测试数据库中包括 50 个说话人(25 男 25 女)，平均每人读 100 个单词和 30 个句子。英语发音脚本分为单词和《新概念英语》中的短文，覆盖了英语中所有的音素。

此外，邀请了 2 位语言学和音系学专家独立地对录制的测试数据库和自适应数据库进行 5 分制评分。然后将自适应数据库中所有专家评分为 5 分的发音作为最终的自适应数据库。

3.2 标准 HMM 的训练

实验采用剑桥大学提供的 HTK 工具包作为 HMM 训练的平台。训练英语母语 HMM 时，采用 TIMIT 建议的 4 620 个句子，语料发音长度总和约为 3 h 49 min，训练中，将 TIMIT 中的 61 个音素映射为 CMU 词典中的 39 个音素。

由于母语为英语的声学模型和母语为非英语的声学模型间存在严重的失配问题，因此在用母语的声学模型识别非母语时，识别率会急剧下降。为了弥补这一缺陷，需要对母语的声学模型进行自适应。在实验中，对母语声学模型先采用 MLLR(Maximum Likelihood Linear Regression)[7]方法进行全局自适应，然后利用MAP(Maximum A-Posteriori)[7]方法进行局部自适应，最终得到具有中国人发音特征的英语标准发音的HMM。

3.3 UBM 的训练

UBM 要能反映出各个音素标准发音间共有的一些特征分布，实验中使用 TIMIT 发音数据库作为训练数据库，训练工具为 HTK。训练时，可将 UBM 当作具有一个状态的 HMM，将 TIMIT 数据库发音脚本中的 39 个音素统一改为“ubm”，取训练好的 HMM 的唯一状态作为 UBM。UBM 的高斯混合数采用 64。

4 实验结果

4.1 标准发音模型的性能

经过训练，最终得到三音子(triphone)HMM 1 872 个，实际存在的状态数为 596 个，每个状态有 8 个对角协方差矩阵的高斯分量。三音子模型的识别率如表 1 所示。

表 1 三音子 HMM 模型的音素识别率 (%)

声学模型	母语	非母语
母语	73.76	36.93
自适应后	-	57.13

从表 1 可知，用母语的声学模型识别母语时，识别率很高，而用母语的声学模型识别非母语时，识别率急剧下降。在对母语声学模型作自适应后，识别率提高了 20.2%。

4.2 专家评分相关性

在计算专家评分之间、专家与机器评分之间的相关性时，为了更好地体现每一个分数段的特性，随机选择一个专家，然后从该专家的评分中，随机选择 1 分~5 分的测试数据，并使 5 个分数段的测试数据量接近相等。以后在讨论专家间的相关性及专家和机器评分的相关性时，均以该测试数据集为准。最终选择的 5 个分数的测试数据分布如下所示。对该测试集进行分析，2 个专家评分间的相关性为 0.722。

分数	1 分	2 分	3 分	4 分	5 分	总和
数量	382	325	312	321	348	1 688

4.3 算法之间比较

在实验室自行采集的非母语语音测试集上，本文算法和其他评分算法与专家评分间的相关性比较结果如下。其中，UBM 代表本文的评价算法，在实验室自行采集的非母语语音测试集上，本文的算法要优于其他算法。

算法	UBM	GOP[3]	对数似然[1]	段时长	语速[2]
相关性	0.662	0.622	0.415	0.268	0.115

4.4 扩展 UBM 算法

高斯混合模型是连续隐马尔科夫模型的一个特例，当连续分布的HMM每个状态的观察概率分布满足高斯分布时，高斯混合模型可以看成单状态的连续分布HMM。图 2 是一个具有 3 个混合数的混合高斯分布模型，图 3 是三状态各状态遍历的连续HMM例子[8]。

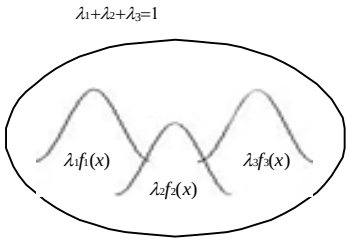

图 2　具有 3 个混合数的高斯混合分布模型

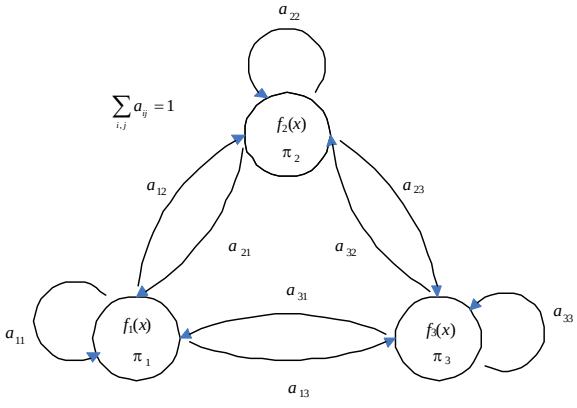

图 3　具有 3 个状态且各状态遍历连续的 HMM

相对于无记忆的 GMM，HMM 能够将各音素发音的不同阶段划分得更加细致，可认为具有 3 个状态 HMM 刻画了音素发音前、发音中和发音后 3 个阶段的与音素无关的特征分布。所以，实验中将 UBM 模型的概念扩展到 HMM 中，将所有音素的发音数据训练成一个具有 3 个状态的 HMM 模型，即扩展 UBM。训练数据库仍然用 TIMIT 发音数据库，训练工具为 HTK，扩展 UBM 状态数为 3，每个状态的高斯混合数为 18。在实验室自行采集的非母语语音测试集上，实验结果如下：

算法	UBM	扩展 UBM
相关性	0.662	0.700

可以看出，扩展 UBM 算法的性能明显优于扩展前的 UBM 算法，主要原因是扩展 UBM 具有 3 个状态，更加细致地刻画了各个音素发音不同阶段的特征分布，实验结果为 0.700，接近专家间的相关性。

5　结束语

本文提出了一种新的发音质量评价算法，将说话人识别领域的 GMM-UBM 模型引入到音素发音质量评价中，训练得到与音素无关的特征分布模型。实验证明，在评分准确性和稳定性方面，该算法明显优于其他评分算法。此外，还将 UBM 的概念扩展到 HMM 中，利用一个具有 3 个状态的 HMM 表征全局背景模型，由于扩展 UBM 具有 3 个状态，更加细致地刻画了各个音素发音不同阶段的特征分布，因此评分效果优于其他算法，接近专家的评分相关性。

该方法还有很多需要继续研究和改进的地方，首先，UBM 模型的训练受数据库数据的影响很大，如何训练一个好的 UBM 模型是研究的重点和难点。其次，在做强制对齐时，发音的不准确会造成音素切割的不精确，造成评分的不准确。将 GMM-UBM 的概念引入到音素发音纠错中也是以后研究的方向。

参考文献

[1] Neumeyer L, Franco H, Digalakis V. Automatic Scoring of Pronunciation Quality[J]. Speech Communication, 2000, 30(2): 83-93.

[2] Franco H, Neumeyer L, Digalakis V. Combination of Machine Scores for Automatic Grading of Pronunciation Quality[J]. Speech Communication, 2000, 30(2): 121-130.

[3] Witt S M, Young S J. Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning[J]. Speech Communication, 2000, 30(2): 95-108.

[4] 梁维谦, 王国梁, 刘　加. 基于音素的发音质量评价算法[J]. 清华大学学报: 自然科学版, 2005, 45(1): 37-45.

[5] 刘振安, 王晋军, 孙　捷. 基于数字串内容识别的用户验证方法研究[J]. 测控技术, 2005, 24(9): 7-14.

[6] Reynolds D A, Quatieri T F, Dunn R B. Speaker Verification Using Adapted Gaussian Mixture Models[J]. Digital Signal Processing, 2000, 10(1-3): 19-41.

[7] Steve Y, Evermann G, Kershaw D. The HTK Book(for HTK Version 3.2)[D]. Cambridge: Engineering Department of Cambridge University, 2002: 134-143.

[8] 赵　力. 语音信号处理[M]. 北京: 机械工业出版社, 2003: 249-251, 256-257.