发音质量自动评测技术

AUTOMATIC EVALUATION TECHNOLOGY OF PRONUNCIATION QUALITY

张 珑 严 可 季伟东 王建华 著 李海峰 主审

哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

发音质量自动评测技术

张　珑　严　可　季伟东　王建华　著
李海峰　主审

哈爾濱工業大學出版社

内容提要

本书结合作者在发音质量自动评测研究领域最新的科技成果，系统地介绍了发音质量自动评测的基本理论、技术和方法。本书首先详细叙述了发音质量自动评测的技术框架、发展历程及存在的挑战；然后从评测特征提取和评测模型训练两个角度，分别介绍与音素相关的后验概率变换、针对发音质量评测的声学模型训练、基于评测性映射变换的无监督声学模型自适应和系统相关的评测性映射变换的训练及统一框架；最后结合汉语普通话音节的三元结构和音韵特点，分别对汉语普通话声韵母、声调和儿化音变的发音质量自动评测技术进行针对性研究和创造性方法改进。

本书可作为高等院校智能信息处理、模式识别、语音信号处理、教育技术等学科高年级本科生和研究生的教学用书，也可作为从事计算机辅助语言学习、口语自动化评测的科技人员了解发音质量自动评测领域动向及应用新技术的重要参考书。

图书在版编目(CIP)数据

发音质量自动评测技术/张珑著. —哈尔滨：
哈尔滨工业大学出版社，2015.6
ISBN 978-7-5603-5415-6

Ⅰ.①发… Ⅱ.①张… Ⅲ.①发音-质量-自动检测
Ⅳ.①H018.1

中国版本图书馆 CIP 数据核字(2015)第 165954 号

策划编辑 王桂芝
责任编辑 郭 然
出版发行 哈尔滨工业大学出版社
社 址 哈尔滨市南岗区复华四道街 10 号 邮编 150006
传 真 0451-86414749
网 址 http://hitpress.hit.edu.cn
印 刷 哈尔滨工业大学印刷厂
开 本 787mm×1092mm 1/16 印张 12 字数 291 千字
版 次 2015 年 6 月第 1 版 2015 年 6 月第 1 次印刷
书 号 ISBN 978-7-5603-5415-6
定 价 48.00 元

(如因印装质量问题影响阅读，我社负责调换)

前　言

发音质量自动评测(Automatic Pronunciation Quality Evaluation,APQE)是计算机辅助语言学习(Computer Assisted Language Learning, CALL)及口语考试中的核心技术问题,其研究成果对提高学习者学习的灵活性和满意度、减少人工阅卷的主观性和不稳定性、降低投入成本、提高实效性具有重要的理论意义和科学价值,应用前景广阔。随着国内普通话的大力推广和普及,以及国外汉语学习热潮的快速兴起,针对汉语普通话的发音质量自动评测技术实际需求强劲,且更具特色和挑战性,有必要深入系统地研究。

汉语是一种单音节声调语言,每个音节包括声母、韵母和声调三部分,音节间界限较分明,有鲜明的轻重音和儿化音。汉语音节的三元结构及音韵特点与英语语音差异较大,需要结合评测任务和汉语语言特点,在表征、建模和计算等方面进行针对性研究和创造性方法改进,以提高声学模型的精度和评测模型的准确度。

本书结合作者在该领域研究的科技成果,系统地介绍了发音质量自动评测的基本理论、技术和方法。本书从两个角度分别探讨母语人群的汉语普通话发音质量自动评测技术:一是从整体评测角度入手,深入探讨在仅有专家篇章级别标注情况下的发音质量自动评测技术(详见本书第2~5章);二是从细节评测角度入手,深入探讨在获取专家音素级别精细标注情况下的发音质量自动评测技术(详见本书第6,7章)。

在整体评测层面,首先针对不同音素后验概率测度不能一致地描述音素发音质量的缺陷,提出一种可训练的与音素相关的后验概率变换方法。通过对不同音素的后验概率进行相应的变换,使得变换后的音素发音质量测度能更加一致地描述发音质量。接下来,针对目前发音质量自动评测系统中声学建模的缺点,提出一种针对评测的声学模型训练算法。该方法利用覆盖各种不同发音质量的数据,通过最小化机器分与人工分均方误差准则进行声学模型训练,因此,可得到与评测目标紧密相连的声学模型(称为评测声学模型)。紧接着,针对评测声学模型难以进行有效的无监督自适应的问题,提出一种利用评测性映射变换(Evaluation-oriented Mapping Transform,EMT)的无监督自适应方法。EMT仍利用覆盖各种不同发音质量的数据,通过最小化机器分与人工分均方误差得到,因此具有与评测目标紧密相连的性质(即"评测性")。在测试时,通过将EMT直接应用至自适应后的声学模型,能将这种评测性"映射"到该声学模型上,得到说话人相关的评测声学模型;然后,考虑评测系统具体应用存在的问题,提出了将具体评测系统融入EMT训练的统一理论框架。利用统一框架能得到更符合具体的评测系统要求的EMT,使系统性能得到进一步提升。在国家普通话水平测试(PSC)现场录音语音库上的实验结果表明(篇章级别上进行整体评测),将音素相关后验概率变换融入EMT训练统一框架中得到了显著超过人工评分一致度的性能,表明该方法能很好地解决后验概率测度的两个问题。

在细节评测层面,针对汉语普通话发音特点和发音规律,以提高人机评分相关性和降低机器评分错误率为目标,模拟人工专家评测的过程,从声韵母、声调、儿化音变三个层面,选取具有代表性的鲁棒评测特征,构建更加精细的声学模型和更加准确的评测模型,用来提升

汉语普通话发音质量自动评测方法的实际性能。针对经典的发音良好度(Goodness of Pronunciation,GOP)算法存在的问题,提出一种基于音素混淆概率矩阵的声韵母发音质量自动评测方法,提高了音素段切分的准确性,同时有效降低声学模型间的相似度,提高计算的精度。针对包含错误发音的数据容易获取,但标注困难、不易利用的问题,提出一种基于扩展发音空间的声韵母发音质量自动评测方法,提高了声学模型的适应性和覆盖范围,同时设计对错误发音数据进行聚类的非监督学习策略,可实现发音质量评测模型的自动更新。针对多层次基频特征的综合利用问题,提出一种基于系统融合的多维置信度的声调发音质量自动评测方法,建立嵌入式和显式混合声调模型,能同时利用长时语段和短时语段的基频特征,且避免了单维置信度分数加阈值判断方式的缺点,有效提高了声调发音质量评测方法的准确性。针对汉语儿化音复杂多变、很难采用传统的评测方法进行有效评测的问题,提出一种基于分类思想的儿化音发音质量自动评测方法。结合儿化音的发音规律和声学特性,优选儿化音的多种代表性特征,包括共振峰、发音置信度、时长等,同时提出了一种改进的AdaBoost集成学习方法,该方法重新设计了基分类器的权值计算方法和迭代更新策略,特别适合数据分布不平衡的多类分类问题,实现了对儿化音发音质量的有效分类,分类效果明显优于 AdaBoost 分类器和其他经典单一分类器。通过综合声韵母、声调和儿化三个方面的评测结果(音节级别上进行细节评测),系统实际评测性能得到很大提升,音节分差下降到4.26,与人工评测的3.71 非常接近,说明机器自动评测可以代替人工评测在大规模语言考试中应用。

本书由张珑、严可、季伟东、王建华共同撰写,具体编写分工为:第1,6,7,8 章由张珑撰写,第2,3,4 章由严可撰写,第5 章由王建华和季伟东共同撰写。全书由张珑统稿,李海峰主审。另外,单琳琳、张鹏、段喜萍、赵云雪等人在本书的图片处理、数据收集方面做了很多工作,在此表示感谢。

由于作者水平有限,书中难免存在不妥之处,望读者批评指正。

作 者

2015 年 3 月

此为试读,需要完整PDF请访问: www.ertongbook.com

目　录

第1章　发音质量自动评测技术概论

1.1　引　　言

发音质量自动评测(Automatic Pronunciation Quality Evaluation, APQE)一般是让说话人朗读给定文本,计算机对其发音进行自动分析,计算出发音质量的置信度,最后反馈出具体等级或者分数[1]。它的目标是赋予计算机担任虚拟教师的能力,对学生的发音质量进行公正、客观、高效的评测,缓解专业口语教师严重稀缺的问题。在学习上,它能帮助学生更好地了解实际发音水平,提高口语学习效率和促进自学的进行;在考试上,它能辅助或者代替人工进行口语考试的阅卷,大幅提升阅卷效率及质量[2,3]。因此,发音质量的自动评测技术日益成为语音信号处理和现代教育技术的研究热点。

发音质量自动评测就其本质而言,是对人工主观评测过程的模拟,并通过对人工评测结果的机器学习,达到甚至超过人类专家的评测性能。为此,本章首先从人类认知的角度入手,分析人工评测的整个过程,给出发音质量自动评测的基本原理和功能结构;接着介绍如何在语音识别技术框架下,搭建出一个基本的发音质量自动评测系统,并对系统中主要功能模块的关键技术和优化方法进行详细探讨;然后介绍研究用语音数据库的采集、录制和人工标注方法,并给出多种系统性能评价方法,用来从不同侧面反映实际应用系统的各项性能指标;最后通过文献综述的方式,对国内外发音质量自动评测技术的发展历程、主要技术方法和实际应用系统进行了详细阐述,并进一步提出了当前研究存在的主要问题和面临的技术挑战。

1.2　发音质量自动评测的基本原理

从认知的角度看,评测专家对待评测语音进行人工评测的过程如图1.1所示,需要经历感觉、知觉、理解评测三个阶段[4]。感觉阶段指人的听觉系统对待评测语音信号进行接收和初步处理,获取其语音表征。知觉阶段是指将上述语音表征识别或者关联为某种语言形式的心理过程(分别对应与文本无关和与文本相关的发音质量评测),其中语言形式是指特定类别语音表征在心理上留下的经验和认知(心理表征),现代神经科学认为,这些语言形式以某种方式存储于大脑之中[5]。由于感觉和知觉很难截然分开,一般被统称为感知。理解评测阶段是对从待评测语音中提取的语音表征,与存储在大脑中的某一固有语言形式所能容许的语音表征进行匹配和比对的过程[6]。当评测专家在感知过程中获得的语音表征与固有语言形式所允许的语音表征相比发生偏离或者越界时,专家将依据此偏移量判断发音质量的标准程度。

发音质量自动评测正是采用一种模型化的方式来模拟人类评测专家的评测过程,类比图1.1,其功能结构如图1.2所示。从图1.2中可以看出,感觉阶段对应发音质量自动评测

中语音信号的特征提取，获得代表性发音特征。知觉阶段的语言形式对应着标准发音单元模型集合（广义上讲，可以是不同粒度上不同类别的多种标准发音单元模型集合，比如音节模型集合、声韵母模型集合、声调模型集合等），感知结果对应着在标准发音单元模型集合上的识别结果或者关联结果（分别对应与文本无关和与文本相关的发音质量自动评测）。理解评测阶段的匹配偏差程度对应着发音特征由对应标准发音单元模型生成的置信度，评测结果对应着最后评定的发音质量等级或者分数。

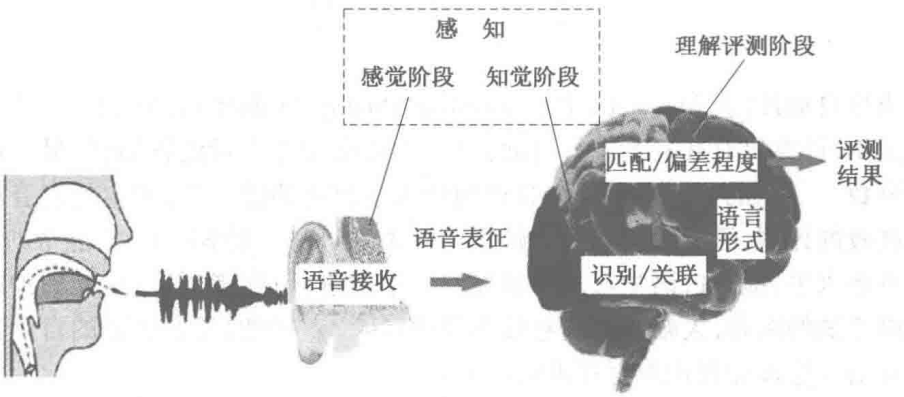

图 1.1 基于认知理论的人工评测过程

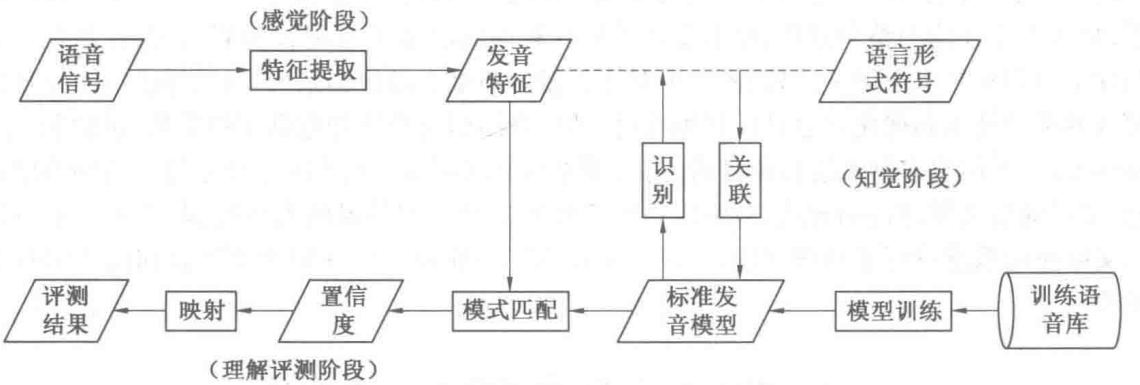

图 1.2 发音质量自动评测的基本功能结构图

从图 1.2 也可以看出，发音质量自动评测任务需要解决的主要问题是，待评测语音与标准发音单元模型之间的匹配程度问题。因此，发音质量自动评测可以转化为计算语音识别中待评测语音段能够被正确识别为其对应的标准发音单元模型的置信度问题（即信号 X 被解码成模式 P 的置信度）。基于这样的思想，发音质量自动评测的目标是寻找适合的评测特征或者特征集，使得这些特征对于发音标准的语音可以得到较高的分数，而对于发音不标准的语音则得到偏低的分数。因此，可以借鉴语音识别技术的基本框架来进行发音质量的自动评测。

1.3　发音质量评测系统的基本功能

1.3.1　系统结构框架

根据 1.2 节发音质量自动评测的基本功能结构，基线系统的结构框架如图 1.3 所示，其中核心功能包括发音特征提取、发音模型训练及发音质量评测，在下面几个小节中将对它们进行详细介绍。

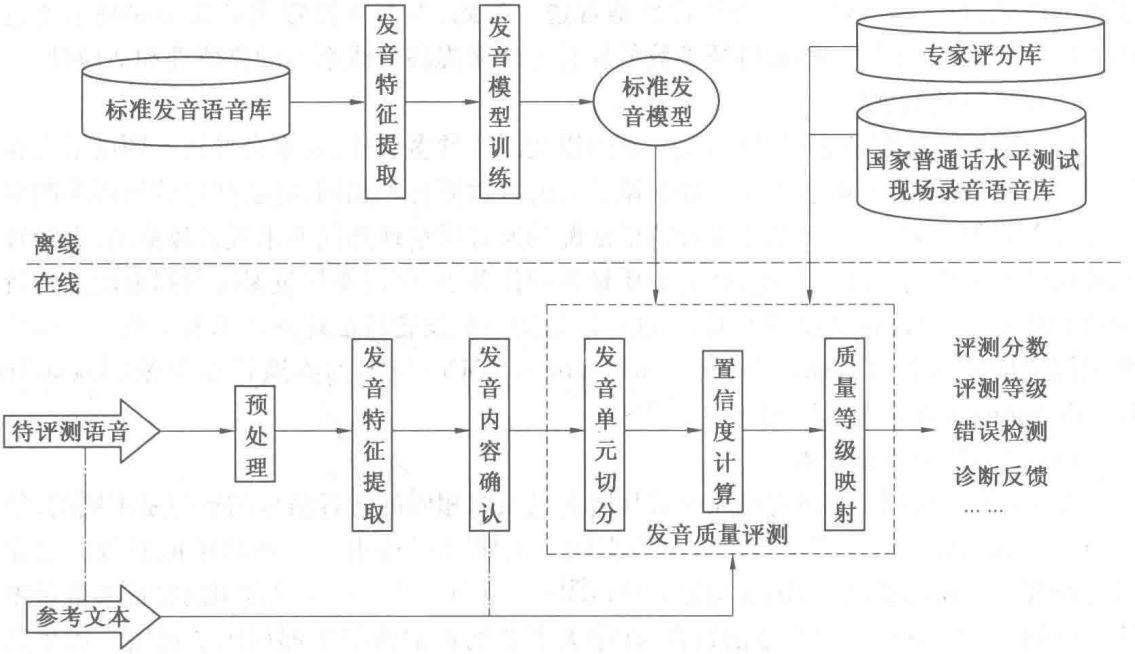

图 1.3　发音质量自动评测基线系统的结构框架图

预处理：为了保证后续处理的鲁棒性，在实用系统中，一般需要对输入语音进行一些预处理，比如去除噪声、活动语音检测、语音增强等，以便进行准确的语音识别及发音质量自动评测。

发音内容确认：针对发音质量自动评测任务而言，由于待评测人是按照参考文本进行发音的，因此待评测语音的对应文本是已知的。而对于胡乱发音或者故意错读的语音没有进行发音质量评测的必要和意义，通过发音内容确认模块，只对那些基本符合参考文本，且基本完整的发音进行评测，能有效地提高评测系统的鲁棒性。发音内容确认可参见文献[7，8]，本书不做具体研究。

发音质量评测：首先根据已知参考文本，利用标准语音识别器将待评测语音按照基本发音单元进行切分；第二步提取切分后识别的似然度数值（置信度计算）；第三步把似然度数值映射为专家评测等级或分数（发音评分），或者将似然度数值符合某阈值范围的发音判为发音缺陷或者错误（错音检测）。

1.3.2　发音特征提取

1. 短时谱特征

语音信号一般被看作是短时平稳信号，因此在处理语音信号时，常常需要对语音信号做

分帧处理,然后提取每个语音帧的特征参数。常见的特征参数有梅尔倒谱系数(Mel Frequency Cepstrum Coefficient,MFCC)、线性预测倒谱(Linear Predictive Cepstrum,LPC)系数、感知线性预测(Perceptual Linear Prediction,PLP)系数、线谱对(Line Spectrum Pair,LSP)参数等。上述谱特征都具有很强的通用性,在各类语音相关任务中都得到了广泛的应用,比如语音信号处理、语音识别、说话人识别等。一些对比研究结果表明,对谱特征本身的鲁棒性处理常常比选择哪种谱特征类型更重要[9]。不同类型的谱特征之间具有一定的互补性,不同类型的谱特征的融合对系统性能的提升有一定帮助[10]。因此,在发音质量自动评测领域,很少提出有针对性的新的谱特征类型,大多数研究者都是直接利用已有的谱特征或者进行细微的改进,且一般 MFCC 和 PLP 特征是首选。为此,本书中发音质量自动评测系统也采用 MFCC 特征,并采用一些谱特征参数的规整方法来提高基线系统的鲁棒性和适应性。

2. 特征参数的规整

由于应用环境复杂多样,说话人的不同以及语音的多变性,常常会导致不同说话人在发同一音素时,所呈现的声学现象会随着说话人的特性而各不相同,这必然导致所提取的发音特征也会有很大差异,这对基于发音特征匹配的发音质量评测任务来说是致命的,会导致评测系统性能的严重下降。因此,在发音质量评测任务中,更需要尽量保证所提取的发音特征与录音环境、录音设备及说话人无关,这可以借助一些发音特征规整技术来实现。下面分别介绍倒谱均值规整(Cepstral Mean Normalisation, CMN)技术和声道长度规整(Vocal Tract Length Normalization,VTLN)技术。

(1)倒谱均值规整技术。

在不同的环境下,不同麦克风或者相同麦克风对相同的语音信号的响应是不同的,传输信道会对输入语音信号产生卷积噪声的影响。文献[11]提出了一种简单但有效的去除信号卷积噪声影响的算法,即倒谱均值规整(CMN)。CMN 是一种非常常用的特征参数规整方法,它具有简单、鲁棒而且有效的特性,在绝大多数的识别系统中都得到了运用。该算法的具体做法是对训练集和测试集中的每个语音段的语音信号都进行变换,变换方法是在其倒谱特征的基础上减去其所在语音段的倒谱特征均值,这样变换前后信号的概率密度分布相同,只相差一个常数,相当于在横轴上做一个平移,而变换后信号倒谱均值为零,一方面可以补偿说话人之间的差异,另一方面也可以消除环境和信道中的噪声,有助于提高特征的鲁棒性。该算法的不足之处在于,当语音信号比较短时(小于 2 s),比如对只包含一个音素的语音段,其倒谱均值可能估计不准,此时进行 CMN 处理会导致错误率增加;而且倒谱均值本身对区分不同音素是有信息量的,此时进行 CMN 处理会降低特征区分性。所以,一般需要在较长语音信号上进行 CMN 会更为有效和准确。

(2)声道长度规整技术。

说话人对声学特征的影响非常复杂,它不仅来源于说话人生理上的差异,还来源于说话人语言特点的差异。但是,研究者普遍认为:声道的形状,特别是声道的长度(Vocal Tract Length,VTL)是人与人之间发音不同的主要影响因素。因此,如果能把不同说话人的声道长度规整到一个标准长度,即进行声道长度规整(VTLN),能够有效消除说话人的不同。VTLN 方法主要是对说话人的频谱做变换,包括线性变换和非线性变换两种,通过将不同说话人的共振峰规整到近似的位置,以减少声学特征咱不同说话人上的差异。由于采取线性变换和采取非线性变换的方法对语音识别任务中的识别效果差不多,本书选用线性变换的方法,下

此为试读,需要完整PDF请访问: www.ertongbook.com

面进行简要介绍。

首先，简化声道传输模型。假设声道是一段截面均匀的管子，此时声道长度与共振峰成反比，声道长度规整可简单地通过在频域内进行线性变换来实现，计算公式为

$$f' = \alpha f \tag{1.1}$$

式中　f, f'——变换前后的频域变量；

α——频率规整因子。

但是这种简单的直接线性变换会造成带宽的扩大或缩小，进而导致部分频率信息丢失。因此，可采取一种能保持带宽的分段线性变换的方法，其计算公式为式(1.2)，用图形表示如图 1.4 所示。

$$f' = \begin{cases} Af + B & 0 \leqslant f \leqslant F_{\mathrm{L}} \\ \alpha f + \beta & F_{\mathrm{L}} \leqslant f \leqslant F_{\mathrm{U}} \\ Cf + D & F_{\mathrm{U}} \leqslant f \leqslant F_{\max} \end{cases} \tag{1.2}$$

式中　$F_{\mathrm{L}}, F_{\mathrm{U}}$——规整的下边界频率和上边界频率；

$A, B, C, D, \alpha, \beta$——变换参数，由 α, F_{L} 和 F_{U} 确定；

$F_{\max}$——频率的最大值。

式(1.1)、式(1.2)中的最重要的参数就是规整因子 α，一般取值范围设定为 0.8 ~ 1.2，需要提前进行估算确定，方法有基于特征的方法和基于模型的方法。基于特征的方法是通过统计语音的频率特性，直接估计出每个说话人的规整因子，速度快但不稳定。本书采用基于模型的方法，建立隐马尔科夫模型(Hidden Markov Model，HMM)，采用最大似然准则来估计规整因子，相对比较稳定，具体方法可参见文献[12]。

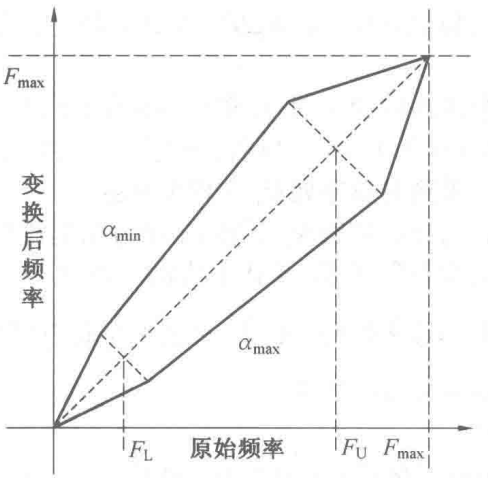

图 1.4　线性频率变换示意图

1.3.3　发音模型训练

1. HMM 模型

隐马尔科夫模型(HMM)是语音识别中声学建模的常用方法，特别适合处理短时平稳的语音信号。HMM 是一种在马尔科夫(Markov)链基础上发展起来的统计模型，它反映着一个双重随机过程，其中一个过程用于描述状态之间的转移，而另一个过程用于描述由状态生

成观测向量。由于观测向量序列是可见的，而状态序列是不可见的，且观测向量与状态之间不是一一对应关系，需要通过观测向量感知到状态的性质及其特性，因此被称为隐马尔科夫模型。

假定一个观测向量序列 $O_T=\boldsymbol{o}_1\boldsymbol{o}_2\cdots\boldsymbol{o}_i\cdots\boldsymbol{o}_{T-1}\boldsymbol{o}_T$ 是符合马尔科夫过程的一个随机变量，它的概率是

$$\Pr①(O_T)=\Pr(\boldsymbol{o}_1)\times\prod_{t=2}^{T}\Pr(\boldsymbol{o}_t\mid O_{t-1})=\Pr(\boldsymbol{o}_1)\times\prod_{t=2}^{T}\Pr(\boldsymbol{o}_t\mid \boldsymbol{o}_{t-1})$$

每个 HMM 中包含若干个状态，记作 $S=\{s_1,s_2,\cdots,s_N\}$，N 为状态数目。那么，O_T 对应的隐状态序列为

$$Z_T=z_1z_2\cdots z_i\cdots z_{T-1}z_T$$

其中，z_t 为 t 时刻所处的状态，$z_t\in S$。

一个 HMM 可以用一个三元组来描述，记作 $\theta\triangleq(\pi,\boldsymbol{A},B)$。

其中，$\pi=(\pi_{s_1},\pi_{s_2},\cdots,\pi_{s_N})$ 为初始状态概率，其中 π_{s_i}（简记为 π_i）表示初始选取的状态为 s_i 的概率，且 $\sum_{i=1}^{N}\pi_i=1$。

$\boldsymbol{A}=[a_{s_i,s_j}]_{N\times N}$ 为状态转移矩阵，其中 a_{s_i,s_j}（简记为 $a_{i,j}$）表示从状态 s_i 转移到状态 s_j 的概率，且 $\sum_{j=1}^{N}a_{ij}=1,1\leqslant i\leqslant N$。

$B=\{b_{s_i}(\boldsymbol{o}_t)\}(1\leqslant i\leqslant N)$ 为输出概率分布，其中 $b_{s_i}(\boldsymbol{o}_t)$（简记为 $b_i(\boldsymbol{o}_t)$）表示在 t 时刻，且处于 s_i 状态时，产生观测向量 $\boldsymbol{o}_t$ 的输出概率。

若 $t-1$ 时刻，其状态为 s_i，则 $\Pr(\boldsymbol{o}_t\mid\boldsymbol{o}_{t-1})=\prod_{j=1}^{N}a_{ij}b_j(\boldsymbol{o}_t)$，且 $\Pr(\boldsymbol{o}_1)=\sum_{j=1}^{N}\pi_jb_j(\boldsymbol{o}_1)$。

对于观测向量序列 O_T，模型 $\theta\triangleq(\pi,\boldsymbol{A},B)$，对应的隐状态序列 Z_T，HMM 存在以下三个最重要的基本问题。

① 评价问题：给定一个观测向量序列 O_T 和模型 $\theta\triangleq(\pi,\boldsymbol{A},B)$，如何计算由该模型产生该观测向量序列的概率，即 $\Pr(O_T\mid\theta)$。目前，解决该问题的主流方法是前向后向算法。

② 解码问题：给定一个观测向量序列 O_T 和模型 $\theta\triangleq(\pi,\boldsymbol{A},B)$，如何获得产生该观测向量序列的最佳状态序列 Z_T。目前，解决该问题的主流方法是维特比（Viterbi）算法。

③ 学习问题：给定一组状态序列 Z_T 及其生成的一组观测向量序列 O_T，训练模型 $\theta\triangleq(\pi,\boldsymbol{A},B)$，并通过调整模型参数 $\tilde{\theta}\triangleq(\tilde{\pi},\tilde{\boldsymbol{A}},\tilde{B})$ 来最大化联合概率 $\prod_{o}\Pr(o\mid\tilde{\theta})$。目前，解决该问题的主流方法是 Baum - Welch 算法。

2. HMM 常用拓扑结构

HMM 一般采用自左向右无跨越的拓扑结构，如图 1.5 所示。

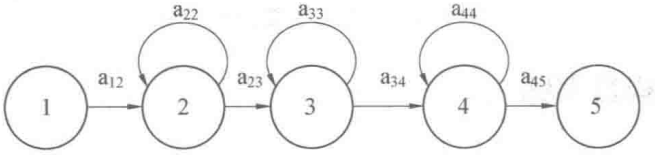

图 1.5 常用的 HMM 的拓扑结构

① 本书中概率（Probability）采用符号 Pr（正体）表示，特此说明。

首尾两个状态分别称为入口状态(Entry State)和退出状态(Exit State),不产生观测向量;其他状态均产生观测向量,因此也称为释放状态①(Emitting State)。例如,我们采用图 1.5的 HMM 描述音素〔a〕(啊)的发音。假设状态 2 表示张口阶段,状态 3 表示发音阶段,状态 4 表示发音结束阶段②。由于人们发音有较强的顺序性,因此发音〔a〕的时候不可能先闭口再张口,也不可能多次张口、闭口。可见,自左向右无跳转的 HMM 能很好地描述人类的发音特性。在这种结构下,无需估计初始概率分布 π。因此,我们需要估计转移矩阵 $\boldsymbol{A}$ 及观测向量的概率分布 B。其中,转移矩阵 $\boldsymbol{A}$ 的估计不在本书讨论范围,下面将简单介绍状态的概率分布 B 的建模及参数估计。

3. HMM 的状态建模及参数估计

虽然语音信号是非平稳信号,但研究表明,在短时间内语音信号可看成是平稳信号。在 HMM 假设中,同一个状态所释放的观测量是平稳的,描述着信号的短时平稳性;而状态之间是不平稳的,描述了在一个较长的时间内,由于发音状态的变化导致所观测的信号不再平稳。例如上文采用一个含三个有效状态的 HMM 描述音素〔a〕的发音,则该建模方式的基本假设就是音素〔a〕的发音可用张口、发音、结束这三个平稳状态描述。由于我们认为同一状态所产生的观测量是平稳的,因此,在当收集足够多的属于该状态所产生的观测量后,就可以估计出该状态的概率分布 B。

当然,对于不同说话人、录音环境,HMM 的状态内平稳假设并不成立。因此,一种好的建模方法需要满足以下两个条件:第一,能较精确地估计出该状态的概率分布参数 B;第二,可通过少量数据对状态参数 B 进行调整,使它能描述当前说话人在当前录音环境下的统计特性;或者能较精确地描述可变环境及说话人情况下的不变量(即发音状态)。目前在语音信号处理中,常用高斯混合模型(Gaussain Mixture Model,GMM)或者人工神经网络(Artificial Neural Network,ANN)对 HMM 的状态进行建模,分别称为 HMM-GMM 和 HMM-ANN 框架,下面分别加以介绍。

(1)利用 GMM 进行 HMM 状态建模(HMM-GMM 框架)。

GMM 即高斯混合模型,它采用多个不同权重的高斯概率密度函数之和表示描述状态的概率密度函数。GMM 不仅能较精确地描述平稳信号的统计特性,还可以通过少量数据对模型参数进行快速调整(称为说话人自适应),使得更新后的模型(通常称为"说话人相关的声学模型")能较好地描述测试集中新说话人、新录音环境的统计特征,是语音信号处理中主流的声学建模方法,本书中的后续工作也将在此基础上进行。

令符号 θ 表示 HMM,它的第 s 状态号记为 θ_s,第 s 状态的第 k 高斯记为 θ_{sk},则观测向量 $\boldsymbol{o}_t$ 的似然度计算如下:

$$\Pr(\boldsymbol{o}_t \mid \theta_{sk}) = \frac{1}{\sqrt{(2\pi)^{\mathrm{Dim}} \mid \boldsymbol{\Sigma}_{sk} \mid}} \exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{u}_{sk})^{T} \boldsymbol{\Sigma}_{sk}^{-1} (\boldsymbol{o}_t - \boldsymbol{u}_{sk})\right] \tag{1.3}$$

其中,$\boldsymbol{u}_{sk}$,$\boldsymbol{\Sigma}_{sk}$分别为该 HMM 的第 s 状态的第 k 高斯的均值矢量和协方差矩阵(一般为对角阵);Dim 为声学特征(即观测量)的维数。因此,该状态的似然度即为所有高斯似然度的加权和,如下所示:

---

① 在语音信号处理中,一般称为"有效状态",本书也将沿用这种称呼。

② 实际 HMM 的训练中并没有精细至状态标注,因此实际的 HMM 的状态物理意义并不明确。该例仅是为了更形象地阐述这种自左向右无跳转的 HMM 建模的思想。

$$\Pr(\boldsymbol{o}_t \mid \theta_s) = \sum_{k=1}^{K_s} c_{sk} \Pr(\boldsymbol{o}_t \mid \theta_{sk}) \tag{1.4}$$

其中，K_s 为第 s 状态的高斯数目；c_{sk} 为第 s 状态的第 k 个高斯的权重。

(2)利用 ANN 进行 HMM 状态建模(HMM-ANN 框架)。

相比 GMM，ANN 更加擅长描述可变的发音人及录音环境中的不变量，因此也受到广泛重视。利用 ANN 进行 HMM 状态建模的方法最早由 Ikbal 等人提出[13]，但由于 ANN 训练的自由度会随着 MLP 的层数增加而呈指数增长，因此当时的研究仅限于浅层的 ANN，学习能力不强，收益有限。随后，在 2009 年，Y. Bengio 在工作中提出了一种名叫 DBN(Deep Belief Network)可深层学习的 ANN[14]，并逐渐成为统计模式识别的研究热点。随后，微软的研究人员将其应用于语音识别中，在不做自适应的情况下，取得了显著超过 GMM-HMM 的识别性能，并受到语音信号处理研究人员的高度关注[15,16]。

然而，目前 HMM-ANN 或者 HMM-DBN 的研究尚不成熟，人们仍难以利用少量样本进行可靠的声学模型参数调整，因此暂时无法取代 HMM-GMM。另外，在本书所研究的发音质量评测任务中，尚未发现有应用 DBN 取得显著收益的报道。因此，后续的研究工作仍在 HMM-GMM 框架下进行。

4. EM 算法及 HMM 的状态参数的最大似然估计

期望最大化(EM)算法是统计学习中重要的最大似然估计(Maximum Likelihood Estimation，MLE)方法，也是 HMM-GMM 参数估计的基石。同时，EM 也是本书后续章节的针对评测的声学建模的重要理论基础，因此本节将详细介绍 EM 算法及如何利用 EM 算法估计 HMM 中的状态参数。

(1)EM 算法简介。

EM 算法是 HMM 训练问题的中 Baum-Welch 重估计的基石，它解决了在不完全数据下最大似然估计的问题[17,18]。EM 的主要思想是通过引入适当的辅助函数，将不完全数据的最大似然估计转化为完全数据的最大似然估计，并通过对辅助函数的优化，从而达到间接的优化不完全数据的对数似然度的目的。

这里，我们讨论的 EM 算法不局限于语音信号处理，因此采用符号 X 表示观测序列，Y 表示隐含的状态序列，Φ 表示数学模型。我们的目标是根据不完全数据优化目标函数 $\log \Pr(X \mid \Phi)$。

根据贝叶斯公式可知 $\Pr(X, Y \mid \Phi) = \Pr(Y \mid X, \Phi) \Pr(X \mid \Phi)$，两边取对数，有

$$\log \Pr(X \mid \Phi) = \log \Pr(X, Y \mid \Phi) - \log \Pr(Y \mid X, \Phi) \tag{1.5}$$

上式两边对观测序列 X 在旧模型 $\Phi^{(0)}$ 下的隐变量 Y 求期望，可得

$$E_{S \mid X, \Phi^{(0)}}[\log \Pr(X \mid \Phi)] = E_{S \mid X, \Phi^{(0)}}(\log \Pr(X, Y \mid \Phi)) - E_{S \mid X, \Phi^{(0)}}(\log \Pr(Y \mid X, \Phi)) \tag{1.6}$$

令

$$Q(\Phi \mid \Phi^{(0)}) = E_{S \mid X, \Phi^{(0)}}(\log \Pr(X, Y \mid \Phi)) = \sum_{Y} \Pr(Y \mid X, \Phi^{(0)}) \log \Pr(X, Y \mid \Phi) \tag{1.7}$$

$$H(\Phi \mid \Phi^{(0)}) = E_{S \mid X, \Phi^{(0)}}(\log \Pr(Y \mid X, \Phi)) = \sum_{Y} \Pr(Y \mid X, \Phi^{(0)}) \log \Pr(Y \mid X, \Phi) \tag{1.8}$$

于是可得

$$\log \Pr(X \mid \Phi) = E_{S \mid X,\Phi^{(0)}}[\log \Pr(X \mid \Phi)] = Q(\Phi \mid \Phi^{(0)}) - H(\Phi \mid \Phi^{(0)}) \tag{1.9}$$

另外，根据杰森不等式 $\sum_i a_i \log x_i \leqslant \log \sum_i a_i x_i$（其中 $\sum_i a_i = 1$ 且 $a_i \geqslant 0$）有

$$H(\Phi \mid \Phi^{(0)}) - H(\Phi^{(0)} \mid \Phi^{(0)}) = \sum_Y \Pr(Y \mid X,\Phi^{(0)}) \log \frac{\Pr(Y \mid X,\Phi)}{\Pr(Y \mid X,\Phi^{(0)})} \leqslant$$

$$\log \sum_Y \Pr(Y \mid X,\Phi^{(0)}) \frac{\Pr(Y \mid X,\Phi)}{\Pr(Y \mid X,\Phi^{(0)})} = \log \sum_Y \Pr(Y \mid X,\Phi) = 0 \tag{1.10}$$

于是可知：

$$\log \Pr(X \mid \Phi) - \log \Pr(X \mid \Phi^{(0)}) \geqslant Q(\Phi \mid \Phi^{(0)}) - Q(\Phi^{(0)} \mid \Phi^{(0)}) \tag{1.11}$$

上式表明了在每一步迭代过程中，对辅助函数 Q 进行优化的同时，目标函数也会得到优化，且优化幅度比 Q 的优化幅度更大。分析辅助函数 Q，不难发现 Q 分为两部分，其中 $\Pr(Y|X,\Phi^{(0)})$ 为隐状态在更新前模型下的概率，在 HMM 估计中可采用前后项算法（即 HMM 的估计问题）求得；对于第二部分 $\Pr(X,Y|\Phi)$，隐变量 Y 已经不再“隐藏”，因此可直接进行最大似然估计。可见，通过 EM 算法，我们将对复杂的目标函数的优化简化为对较简单的 Q 函数的优化。

（2）利用 EM 算法估计 HMM 的状态（GMM）参数。

对于一个时长为 T 的观测向量序列 $O_T = \boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T$，根据 GMM（假设其为某个 HMM 的第 s 状态，记为 θ_s）计算对数似然度的公式为

$$L(\theta_s) = \sum_{t=1}^{T} \log \Pr(\boldsymbol{o}_t \mid \theta_s) = \sum_{t=1}^{T} \log \sum_{k=1}^{K_s} \frac{c_k}{\sqrt{(2\pi)^{\mathrm{Dim}} |\Sigma_{sk}|}} \exp\left[-\frac{1}{2} (\boldsymbol{o}_t - \boldsymbol{u}_{sk})^{\mathrm{T}} \Sigma_{sk}^{-1} (\boldsymbol{o}_t - \boldsymbol{u}_{sk}) \right] \tag{1.12}$$

优化目的任务是最大化通过调节 GMM 的参数，使得 $L(\theta_s)$ 的期望最大化。然而，我们不知道观测向量属于哪个高斯，难以直接优化式（1.12）。通常做法是采用 EM 算法，引入辅助函数 $Q(\theta_s, \theta_s^{(0)})$，将隐含的高斯通过求期望方式“显示”出来，如下所示：

$$Q(\theta_s, \theta_s^{(0)}) = \sum_{t=1}^{T} \sum_{k=1}^{K_s} \gamma_{sk}^{(0)}(\boldsymbol{o}_t) \log \Pr(\boldsymbol{o}_t \mid \theta_{sk}) \tag{1.13}$$

其中，$\theta_s^{(0)}$ 为更新前的 GMM 模型参数①；θ_s 为更新后的 GMM 参数；$\gamma_{sk}^{(0)}(o_t)$ 为更新前的 HMM 模型的第 s 状态第 k 高斯（$\theta_{sk}^{(0)}$）在观测向量 $\boldsymbol{o}_t$ 时的后验概率（Posterior Probability，PP）。对于第二项的 $\log \Pr(\boldsymbol{o}_t \mid \theta_{sk})$，代入式（1.3），可以发现 log 和高斯似然度计算中的 exp 可以消掉，因此 Q 函数是关于均值矢量 $\boldsymbol{u}$ 的二次函数，开口恒向下，因此很容易求解到 Q 的全局最优解。

同时，可以证明，辅助函数与原函数在原点处一阶导相等，如式（1.14）[19]。这个性质将在本书的第 3，4 章中使用到。

$$\frac{\partial Q(\theta_s, \theta_s^{(0)})}{\partial \theta_s} = \frac{\partial L(\theta_s)}{\partial \theta_s}\bigg|_{\theta_s = \theta_s^{(0)}} \tag{1.14}$$

因此，高斯混合模型的估计可归纳如下。

---

① 本书凡是有上标$^{(0)}$的参数表示该参数是根据更新前的声学模型得到。

E 步：根据更新前的声学模型计算所有训练样本的所有时刻的所有高斯的后验概率 $\gamma_{sk}^{(0)}(\boldsymbol{o}_t)$，如式(1.13)所示。其中 $\gamma_s^{(0)}(\boldsymbol{o}_t)$ 为状态 s 在 t 时刻下的后验概率，可由前后项算法（HMM 的第一个问题的解）或者 Viterbi 解码（HMM 的第二个问题的解）得到。

$$\gamma_{sk}^{(0)}(\boldsymbol{o}_t)=\gamma_s^{(0)}(\boldsymbol{o}_t)\frac{c_{sk}^{(0)}P(\boldsymbol{o}_t\mid\theta_{sk}^{(0)})}{\sum_{l=1}^{K_s}c_{sl}^{(0)}P(\boldsymbol{o}_t\mid\theta_{sl}^{(0)})} \tag{1.15}$$

M 步：更新 GMM 模型参数

$$\boldsymbol{u}_{sk}=\frac{\sum_{t=1}^{T}\gamma_{sk}^{(0)}(\boldsymbol{o}_t)\boldsymbol{o}_t}{\sum_{t=1}^{T}\gamma_s^{(0)}(\boldsymbol{o}_t)} \tag{1.16}$$

$$\Sigma_{sk}=\frac{\sum_{t=1}^{T}\gamma_{sk}^{(0)}(\boldsymbol{o}_t)(\boldsymbol{o}_t-\boldsymbol{u}_{sk})(\boldsymbol{o}_t-\boldsymbol{u}_{sk})^{\mathrm{T}}}{\sum_{t=1}^{T}\gamma_{sk}^{(0)}(\boldsymbol{o}_t)} \tag{1.17}$$

例：音素〔a〕的参数的估计，其中音素〔a〕由一个三个有效状态的自左向右无跳转的 HMM－GMM 描述，如图 1.6 所示。

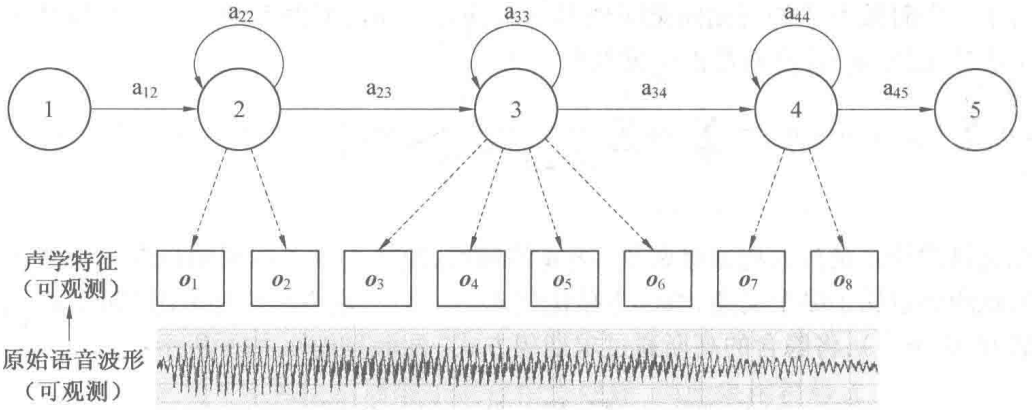

图 1.6 利用 HMM 描述发音〔a〕

E 步：根据更新前的声学模型 $\theta^{(0)}$ 进行语音识别，得到观测向量属于的状态，并计算高斯后验概率。

通过 HMM 的解码（图 1.6），我们可以知道 $\boldsymbol{o}_1$，$\boldsymbol{o}_2$ 由 $\theta^{(0)}$ 的状态 2 所释放，$\boldsymbol{o}_3$，$\boldsymbol{o}_4$，$\boldsymbol{o}_5$，$\boldsymbol{o}_6$ 由状态 3 所释放，$\boldsymbol{o}_7$，$\boldsymbol{o}_8$ 由状态 4 所释放。因此对于观测向量 $\boldsymbol{o}_1$ 而言，状态 2 的后验概率 $\gamma_2^{(0)}(\boldsymbol{o}_1)=1$，其他状态的后验概率（$\gamma_3^{(0)}(\boldsymbol{o}_1)$，$\gamma_4^{(0)}(\boldsymbol{o}_1)$）均为 0，依此类推。对于 $\theta^{(0)}$ 的第 2 状态的第 k 高斯 $\theta_{2,k}^{(0)}$，在状态 2 条件下的高斯后验概率 $\mathrm{PP}(\boldsymbol{o}_1,\theta_{2,k}^{(0)}\mid\theta_2^{(0)})$ 即为该高斯的加权似然度 $c_{2,k}^{(0)}\Pr(\boldsymbol{o}_1\mid\theta_{2,k}^{(0)})$ 除以该状态的所有高斯的加权似然度 $\sum_{l=1}^{K_2}c_{2,l}^{(0)}\Pr(\boldsymbol{o}_1\mid\theta_{2,l}^{(0)})$，即

$$\gamma_{2,k}^{(0)}(\boldsymbol{o}_1)=\gamma_2^{(0)}(\boldsymbol{o}_1)\frac{c_{2,k}^{(0)}\Pr(\boldsymbol{o}_1\mid\theta_{2,k}^{(0)})}{\sum_{l=1}^{K_2}c_{2,l}^{(0)}\Pr(\boldsymbol{o}_1\mid\theta_{2,l}^{(0)})} \tag{1.18}$$

其中，K_2 为状态 2 的高斯数目。可见，E 步并不存在隐变量，所有的高斯后验概率均可根据

此为试读，需要完整PDF请访问：www.ertongbook.com