语音识别技术在外语口语学习中的应用

The Application of Speech Recognition Technology in Foreign Spoken Language Learning

黄骁勇 虞维平 （南京 东南大学远程教育学院 210096）

摘要：语音识别技术应用于外语口语的学习，有力促进了学习者的学习效果和效率。本文就目前语音识别技术在外语口语学习中的应用进行综述，介绍了目前的研究和应用情况，对一些关键技术和问题进行探讨，并对其发展进行展望。

关键词：语音识别 外语口语学习 发音学习 隐马尔可夫模型

1 引言

随着全球一体化趋势的日渐明显，当今社会越来越多的人希望学习和掌握一门或几门纯正流利的外语，以利于更方便地进行交流。然而，由于传统师资的有限和口语练习环境的缺乏，很多学习者的口语水平很难得到有效的提高，“哑巴外语”的现象屡见不鲜。因此，运用 CALL（Computer – Assisted Language Learning）技术为外语口语学习服务，这是一个必然趋势。

口语的学习一般分为两个部分的内容，一是讲述内容包括语法，结构，习惯用语等方面的学习，另一个就是发音的学习，只有正确的发音才能帮助我们正确地表达观点。因此，口语学习主要表现在发音的学习上。而语音技术（尤其是语音识别技术）的不断成熟则为辅助学习者发音的学习提供了可能。

2 研究和应用现状

2.1 研究现状

语音识别技术在 90 年代初开始逐渐成熟，随后一些科研机构开始研究其在外语口语学习领域的应用。其中，美国的 SRI、CMU、IBM，英国的剑桥大学，日本的一些大学以及国内的香港理工大学和清华大学等都在这方面做了较多的研究工作。

目前在这方面的研究主要集中在以下几个方面：

（1）寻找反映发音质量的性能指标，主要集中在对超发音段（指一段发音的语调、重音、语速和韵律等）的研究；

（2）检测和纠正给定的音素级发音错误；

（3）寻找更合理的评分机制；

（4）研究外语口语学习系统的性能评测手段；

（5）针对非母语问题的自适应技术。

2.2 目前的一些应用

目前，语音识别技术在外语口语学习中已经得到了比较广泛的应用，并且出现了不少成熟的成果和产品。表 1 是一些比较典型的基于语音识别技术的外语口语学习系统/工具。

表 1 基于语音识别技术开发的外语口语学习系统/工具

系统/工具	开发机构	支持语言	特点
WebGrader	SRI	法语/英语	基于 Web 的多语言发音练习工具
EduSpeak	SRI	英语、西班牙语等	用于语言教学软件的开发包，支持多种编程接口
Tell Me More	Auralog	德语、汉语等	自动检测和矫正发音错误，基于 3D 动画的视觉反馈
Fluency	CMU	英语	侧重音素和韵律的检测和矫正
LESSON/J	日本岩手大学等	日语	用于 E – learning 的日语口语学习，平台无关
PROnounce English	ReadSay	英语	用于手持设备和嵌入式系统的交互式口语学习系统

3 应用中的问题和关键技术

外语口语学习系统是一个多学科综合的产物，其

具体到语音识别领域而言有一些普遍问题和关键技术需要考虑。下面就对目前外语口语学习系统中使用的这些问题和技术进行一些介绍和讨论。

3.1 语音评分

语音评分技术可以说是基于语音识别技术的外语口语学习系统的最基本和最核心的部分。语音评分技术能够对学习者的发音进行评价，并给予反馈，因而学习者可以通过反馈来检查自己的学习结果。理想的语音评分技术不仅能够对发音段（指一段发音的单个音素，如英语的音素，汉语的声母和韵母等）进行评价，还应该能对超发音段（如英语的重音，汉语的音调），单词的发音，句子中词与词之间的协同发音，句子的语速和流利程度等进行评价。同时，对学习者的发音评价应该具备一致性和稳定性。当前主流的外语口语学习系统基本上都能够实现对发音段的评价，然而能够实现全面评价的系统还很少。

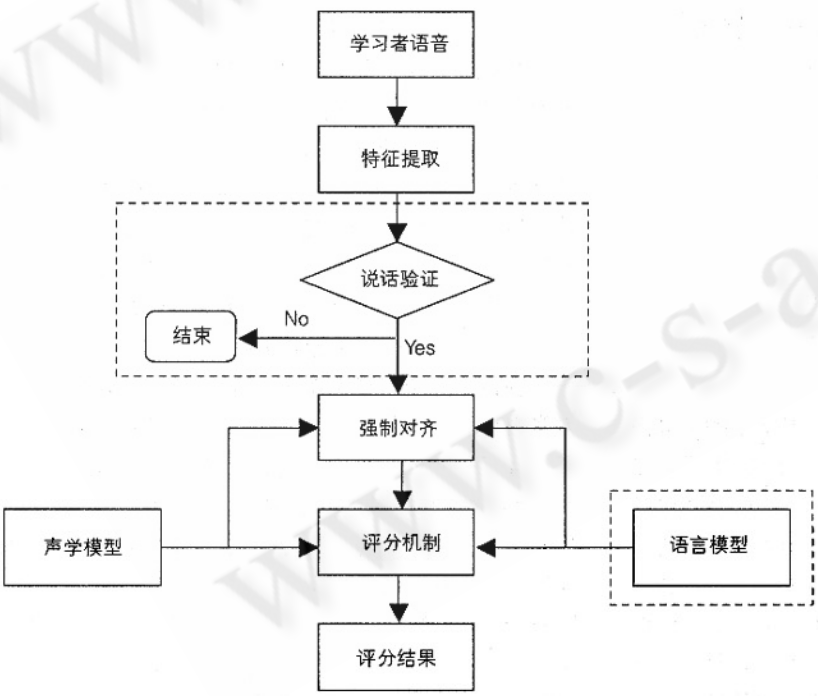

图 1　基于 HMM 的语音评分流程

一般来说，目前的外语口语学习系统按照其所使用的语音评分技术主要可以分为两类。一类是基于语音特征比较的评分方法，它通过对比参考标准语音与学习者的发音，从一个比较主观的角度去评价一段语音的质量，一般采用动态时间规整（Dynamic Time Warping）技术实现，由于其运算量较小，一般运用在手持设备和嵌入式系统中。另一类是基于声学模型（Acoustics Model）的评分方法，通过语音识别技术切割出计算发音质量所需要的小单元（如英语中的音素），再将其与预先训练好的声学模型进行比较并根据评分机制进行评分。这种方式较为客观，主要基于隐马尔可夫模型（Hidden Markov Model）技术实现，目前主流的外语口语学习系统均采用这种技术。下面就以基于 HMM 的语音评分方法为例介绍下要考虑的一些关键技术。

图 1 给出了基于 HMM 的语音评分的流程图，虚线框部分表示可选。如图所示，语音评分的基本环节可以分为以下三步：

（1）首先对学习者输入的语音进行特征提取；

（2）然后以训练好的声学模型（HMM）作为模板，通过强制对齐（通常采用 Viterbi 算法）把语音分割为计算发音质量所需要的小单元；

（3）根据不同的需要（如对音素评分，对单词评分等），采用一定的评分机制进行评分，得出评分结果。另外，在某些情况下，还需要在语音评分的开始加入说话验证（Utterance Verification），以挡下内容和标准发音完全不同的学习者发音，使整个口语学习系统更具可信度。在某些更为复杂的应用中，因为同样的声音可能代表不同的意义，因此需要考虑上下文的关系，以及各种词发生机率的大小，还需要加入语言模型（Language Model），来辅助声学模型做更好的判断。

3.1.1 特征提取

无论是学习者的口语学习还是声学模型的训练，首先要做的第一步都是要提取输入语音的特征参数。特征提取模块的目的是对原始语

音进行处理后,计算语音对应的特征参数,主要包括预处理、分帧、加窗、计算特征参数等过程。

语音识别中常用的特征参数有幅度、能量、过零率、线性预测系数(LPC)、LPC 倒谱系数(LPCC)、线谱对参数(LSP)、短时频谱、共振峰频率、梅尔倒频谱系数(MFCC)等。其中 MFCC 由于其反映了人耳的听觉特征,因而其性能及其鲁棒性是所有参数中最好的。目前多数外语口语学习系统中使用 MFCC 参数来表征语音的内容。然而为了全面反映音强、音调与韵律等的变化,一般还需要提取其他的特征系数来辅助语音的评分,如利用基频轨迹(Pitch Contour)来表征声音音高的变化。

在特征参数提取中,由于每个人声调的高低不同或是录音环境的差异等因素,造成特征参数或多或少会受到一些影响。因此,在提取特征参数之后,还需要对特征参数做一定的规整化(Normalization),使得存在个体差异的特征参数可以在同一基准下进行比较。通常使用的方法有内插法(Interpolation)、线性缩放(Linear Scaling)、线性平移(Linear Shifting)、语者正规化(Vocal Tract Length Normalization)等。

3.1.2 语音对齐

在特征提取之后,要想基于参考模板(声学模型)计算发音质量,必须先将学习者语音的基本评分单元与参考模板进行强制对齐。

这里的参考模板需要通过语料库(Corpus)训练得到,语料的内容取决于系统采用的基本发音处理单元的形式,可以是音素、音节或者单词。声学模型的建立目前一般基于 HMM(也有一些使用 DTW、ANN 等方式)。HMM 采用统计方式来描述语音的特征。在 HMM 方法中,语音序列被看作是马尔可夫(Morkov)随机过程的输出。如果描述了这个 Morkov 随机过程的参数,也就描述了这个 Morkov 随机过程所对应的语音序列。在系统中,主要是利用 HMM 来判断进来的学习者语音的评分单元跟哪个参考模板最相近。然而,我们虽有语音特征的时间序列,但我们并不知道哪个序列要对应到语音的哪个状态(即所谓"隐"马尔可夫模型),所以还需要利用维特比(Viterbi)算法,来找到这两者对应的最佳方式,以使学习者语音中的基本评分单元与参考模板的尽量一致,确保后续评分的准确性。实际应用中在采用 Viterbi 算法进行强制对齐时,一般以音素为基本处理单位,这样可以扩展到与文本无关的情况。

3.1.3 评分算法与机制

发音质量的评价方法以及寻找一个全面衡量学习者发音的评分机制是外语口语学习系统的关键技术之一。

目前已经有不少衡量发音质量的方法。如对发音段进行评价的方法有对数似然度评分(Log likelihood scores)和对数后验概率评分(Log – posterior probability scores)等。对超发音段进行评价的方法有段时长评分(Segment duration scores)和语速评分(Timing scores)等。

显然,仅仅用其中的某一种方法并不能全面客观地评价一个学习者的发音水平。如对数后验概率虽然对学习者发音的内容评价具有很好的稳健性,能够很好地反映学习者的发音与标准发音之间的相似性,但是并不能出反映学习者发音的流利程度和自然度,因此我们还需要考虑一定的评分机制,即综合考虑几种方法来科学合理地评价学习者的发音。这里的评分机制是几种评分方法参数之间的一种线性或非线性关系。一个科学合理的评分机制,应该能让系统模仿语言专家的角色,即系统的评分应该与专家的人工评分保持较好的一致性和稳定性。为了达到这个目的,需要对各评分方法所占的权值进行大量地调整和试验,找出机器评分和人工评分的一个最佳映射关系,才能得到一个比较理想的评分机制。

在得出一个对学习者发音的评分之后,这个得分对学习者来说往往太抽象,与人的感知不相一致,因此还需要基于发音专家知识将这个得分通过映射(通常是非线性的)转换成比较模糊的分类,如"很好"、"好"、"一般"等等,这样比较符合人类的感知习惯,还具有一定的稳定性。但如何使这种得分反馈更有效,更宜于接受、掌握与改进,需要考虑语言声学和心理学等多方面知识。

3.2 非母语问题及自适应技术

一般来说,语音识别系统是针对母语的发音者而设计的。其中的大多数系统采用的是统计模式匹配的技术,系统参数是通过一组代表母语发音者统计特征的数据集合(即标准语料)来训练得到的。相反的,在外语口语学习中,使用系统的学习者却有着不同的学

习背景和特点，而且大多数都为非母语发音者（Non－Native），他们的发音不可能是“纯正”的，或多或少会带有一些“方言”，如法国人在讲英语时总带有一些法语的味道。

显然，在用母语语音训练得到的声学模型与一般学习者的非母语语音之间存在着严重的失配问题，系统的实际性能会受到很大的影响。因此，实际应用中一般还需要对系统模型进行改进。改进系统模型的方法主要有两类：

3.2.1 采用非母语语音去训练声学模型

为了使系统能对各种“方言”在一定的程度上进行“容忍”，因此，在训练声学模型时，除针对母语语料进行的训练之外，还应该同时考虑用非母语语料去训练，这里的非母语语料应该具有广泛的代表性，充分考虑学习者的口音、年龄和性别等特征分布。因此，需要大量的语料，训练过程的代价相对较高。

3.2.2 对现在的声学模型进行非母语自适应

采用非母语语音去训练声学模型，显然需要庞大的语料库的支持，因此，对现在的声学模型进行自适应是一种比较好的选择。这里的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。例如，在学习者使用外语口语学习系统之前，要求其先朗读一篇给出的文章，系统对该学习者的口音进行自适应，并建立相应的模型。这个方法利用系统使用者的少量训练语音，调整系统的参数，提高系统对于该使用者的性能，但是目前尚没有看到这样的自适应能力所要求新用户训练的次数上的一个评价。

3.3 发音错误判断和矫正

一个外语口语学习系统的更高级形式就是能够基于发音专家知识，判断发音错误的类型并给出相应的矫正建议。作为语言的学习者，尽管可以得知系统对其发音质量的评价，但是仅凭这个非母语学习者还是不能了解自己的错误所在，以及该如何来改进这个错误。因此，系统需要帮助学习者检测并定位发音的错误，并对如何改进这些错误给出相应的矫正建议。

3.3.1 检测并定位发音错误

造成错误发音的原因很多，如学生不能体会两种声音的差别或受到其他语言拼读方法的影响，也可能是不会发某种声音等。

检测发音错误的一种做法是是采用语音识别器来检测的。如用母语训练的语音识别器，则对学习者发音的识别错误可以作为发音错误。然而，在没有使用自适应技术的情况下，这样做对非母语学习者的发音很难获得理想的检测结果。目前比较合理的一种做法就是：首先依据发音专家的知识，对发音中容易出现的错误进行分类；然后针对不同的错误类型设计相应的检测算法；最后对学习者发音用各种错误检测算法分别进行检测。

3.3.2 对发音错误的矫正

检测出发音错误之后，依据专家关于发音错误的知识给出错误提示和发音矫正的建议。这些发音专家知识的构建需要经过大量的经验参数的积累，比如大量学习者的发音特征，结合数据挖掘的聚类算法得出不同发音特点的聚类，再经过专家对各类的发音评判，给出各聚类的改进建议。

4 结束语

本文对语音识别技术在外语口语学习中的应用进行综述，介绍了当前的研究和应用情况，探讨了语音识别技术在外语口语学习应用中需要考虑的问题和关键技术。

参考文献

1 蔡莲红、黄德智、蔡锐，现代语音技术基础及应用[M]，北京 清华大学出版社，2003。

2 韩纪庆、王欢良、李海峰等，基于语音识别的发音学习技术[J]，电声技术，2004，(9)：47－51。

3 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，2004，40（1）：112－114。

4 梁维谦、王国梁、刘加 等，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)：5－8。

5 Farzad Ehsani，Eva Knodt. Speech Technology in Computer－Aided Language learning：Strengths and Limitations of a New CALL Paradigm[J]. Language learning&Technology，1998，2(1)：45－60.

6 Kathleen B. Egan . Speaking：A Critical Skill and a Challenge[J]. CALICO Journal，1999，16(3)：277－293.