智能语音技术提升口语教学效率

方强　李爱军　王士进

语言的学习包含了词汇、语法、口语学习等多个方面。从交流的角度看，口语学习恐怕是最关键的。传统的口语教学要求老师发音标准，能实时地对学习者的发音进行评价，纠正学习者错误发音。但在实际的教学过程中，教师少、学生多、课堂时间有限，教师很难在课上对学生进行一对一的口语指导和问题反馈，在一定程度上影响了部分学生的学习效率和参与课堂互动的积极性。如何运用各种现代技术手段，有效缓解口语教学中的痛点，弥补课堂教学的不足，是一件有实践价值和社会意义的事情。随着统计机器学习和深度神经网络的发展，语音技术在语音合成、语音识别、发音逆分析等关键领域有长足进步，使得运用语音技术解决口语学习中的痛点问题成为可能。

1. 语音合成技术可缓解师资问题

语音合成技术是一种将输入的文本转换成语音的技术。传统的语音合成系统通常包括前端和后端两个模块。前端模块主要是对输入文本进行分析，提取后端模块所需要的语言学信息。对于中文合成系统而言，前端模块一般包含文本正则化、分词、词性预测、多音字消歧、韵律预测等子模块。后端模块根据前端分析结果，通过一定的方法生成语音波形。

在实践中，人们通过主观或客观的方法评价语音质量。主观方法就是用实验被试对语音进行打分，比如平均意见分方法（Mean Opinion Score，MOS）、判断韵字测试（Diagnostic Rhyme Test，DRT）、失真平均意见分（Degradation Mean Opinion Score，DMOS）、判断满意度测试方法（Dignostic Acceptability Measure，DAM）。客观方法即是用算法评测语音质量。我们经常会在语音合成的论文中看到通过计算合成语音与真实语音的梅尔倒谱距离（Mel Cepstral Distortion）对合成语音的质量进行评价。但是，目前我们还无法建立一个能完全模仿人类音质感知过程的客观评价系统，只能根据所获得的信息做出尽可能正确的评价，所建立的客观评价模型也与人类感知的评价能力相去甚远。在实际运用中，通常将主观评价和客观评价结合使用，客观评价常用于系统的设计、调整以及现场实时监控阶段，而主观评价则作为实际效果的最终检验。

目前，主流的语音合成系统 MOS 得分都能达到 4.0 以上，能够合成足够标准和自然的语音，用于口语学习中的发音示例，有效缓解口语学习中教师发音水平参差不齐、教师资源严重不足的问题。

2. 语音评测技术仍有薄弱环节

通俗来说，语音评测技术就是一种通过计算机算法自动对发音进行评分和检错并反馈的技术，是计算机辅助语言学习和测试领域中最重要的技术之一，在语言学习和口语考试中扮演着重要的角色。语音评测技术的目标是替代专家和老师，对学习者的发音进行实时自动评估和检错，弥补人工评测主观性强、效率低等方面的不足。

语音评测包括朗读评测和口头表达评测两项关键技术。前者主要包括如字词句篇的朗读题型，考察重点是学习者的发音错误情况和发音质量；后者主要包括如口头翻译、口头复述、看图说话、话题表述等题型，主要考察学习者的逻辑思维能力和语言组织能力。朗读评测技术研究较早，目前已达到成熟水平。口头表达评测技术难度远大于朗读评测，直接使用音频评价学习者的口头表达能力极其困难。

3. 发音逆分析技术尚存在亟待解决的问题

发音逆分析技术是一种从语音声学信号推断发音器官的位置和形状的技术。它和发音可视化技术结合在一起扮演教师的角色，为学习者提供实时视觉的反馈和发音指导。从语音信号逆推发音器官的形状和位置是一项十分困难的工作。不管是基于发音模型的研究，还是基于实际发音数据的研究，都发现有些语音的声学信号与发音器官的位置和形态之间存在一对多的关系，即不同的声道形态能够产生声学特征相似的语音信号。

近年来，随着深度神经网络技术的发展和同步语音—发音数据采集手段的进步，人们将基于双向长短时记忆单元的递归神经网络应用于发音逆推的工作，取得了平均意义上的较好的性能（发音器官的平均位置误差从 2 毫米左右降低到 0.5 毫米左右）。但是，发音逆推技术要应用于口语学习还有一些问题亟待解决。首先，对于每一个具体的发音，现有的发音逆推技术得到的发音器官的形状和位置能否保持发音的音位特征不变，还需要进一步检验。其次，目前大部分发音逆推是基于某一个特定发音人的发音—语音数据开展的，如何将基于特定发音人的发音逆推模型适用于非特定发音人，还是一个尚需深入探索的课题。

4. 发音可视化技术仍处在探索阶段

发音可视化技术可理解为一种结合发音逆分析、口语评测等的结果，将发音人发音器官的位置和形态、发音过程中的气流状态、发音过程中的声带振动状态等信息以视频的形式呈现的技术。发音可视化涉及发音器官的形态建模、发音器官的驱动与展示、气流状态分析与展示、基频曲线的分析与展示等方面。

发音器官的建模及发音器官的驱动与展示等技术主要用于与发音逆分析模型结合，动态展示发音人发音时发音器官的状态，为学习者提供正确的发音动作示例和准确的发音指导。这方面的工作已经开展得比较深入，国内外多家研究单位提出了基于不同数据集的发音器官模型及发音器官运动的驱动模型。但是，现有的发音器官模型和发音运动模型还需要与发音的空气动力学模型结合，进一步验证可视化的发音模型能否准确实现不同音位的语音。

气流状态的分析和展示技术主要用于对发音过程中口腔中气流的状态进行可视化，帮助学习者正确地掌握不同的发音方式。但是，这方面的工作目前还鲜有报道。

基频曲线的分析与展示技术主要用于对学习者口语的韵律特征进行分析和反馈，帮助学习者正确掌握语调，这方面的工作还处于起步阶段。

5. 未来应用前景广阔

如今，应用在汉语和英语口语教学中的智能语音技术日益成熟和稳定，但目前仅集中在口语错误的定位和检测，提供的反馈在形式和内容上都比较单一。对学习者的发音问题，尤其是在语音和语调等方面的问题，缺少直观的视觉反馈来辅助其有效地纠正错误，导致学习者虽然知道问题的存在，但由于缺少可以模仿和参照的发音示范，不知道该从何着手才能克服语音问题、提高发音的准确性。因此，探索和完善可视化反馈是智能语音技术在口语教学应用中的一个重要的研究方向。

随着线下教学模式转向大规模线上教学，人们的学习方式正在发生巨大的变化。结合人工智能技术的语言学习，为在线语言教学打开了一扇大门。实现智能语音技术与课堂教学实践的有机结合，探索科学有效的在线教学方法，也是对新时代语言教学的一个迫切要求。

（本文系国家社科基金重大项目“中国方言区英语学习者语音习得机制的跨学科研究”（15ZDB103）阶段性成果）

（作者单位：中国社会科学院语言研究所；科大讯飞 AI 研究院）

［本文刊于《中国社会科学网-中国社会科学报》2020 年 12 月 22 日］