密级:（涉密论文填写密级，公开论文不填写）

中国科学院研究生院

硕士学位论文

英语发音自动评测技术的研究

作者姓名：陈 蒙

指导教师: 蒙美玲、王岚

中国科学院深圳先进技术研究院

学位类别: 工学硕士

学科专业: 计算机应用技术

培养单位: 中国科学院深圳先进技术研究院

2012 年 5 月

Research on Automatic Scoring of Pronunciation in Oral English Test

By

Meng Chen

A Dissertation Submitted to
Graduate University of Chinese Academy of Sciences
In partial fulfillment of the requirement
For the degree of
Master of Computer Application Technology

Shenzhen Institutes of Advanced Technology,

Chinese Academy of Sciences

May, 2012

致　谢

时间：2008 年 9 月 17 日 10:01（星期三）发件人：陈蒙 收件人：蒙美玲教授

主题：申请做您的研究生

这封邮件是我中科院先进技术研究院生活的开始，是我在环绕智能研究实验室求学的开始，也是我与语音识别这个领域缘分的开始。从大四下半学期来实验室做本科毕业设计算起，研究生生活持续了 3 年半。很庆幸自己在这段有限的时间里有机会认识蒙美玲教授、王岚研究员、罗德安博士以及实验室的兄弟姐妹们，共同度过一段美好而难忘的时光。

感谢蒙美玲教授，虽然与您的接触次数并不多，却能真切的感受到您严谨的治学态度和谦和的为人品德。感谢您当初给予了一位大三的保研生一次珍贵的深造机会，让他有机会接触神奇的语音领域，有机会体验多彩的科研生活。

感谢王岚研究员，感谢您对一个二十几岁的小伙子的信任和器重，让他大胆地去思考、去实践。感谢您周六的清晨给我耐心地改论文，字里行间，我深深地体会到了学术的严谨；还记得艾尔斯系统育才中学推广活动么，从您的言行举止中，我学会了如何在困难面前保持冷静和果敢。您不仅仅教会了我如何研究，还培养了我很多优秀的品质，这些，都是我一生的财富。

感谢罗德安博士，你手把手教我如何查询文献，分析实验结果，甚至是耐心地帮我调试代码。你让我明白了科研工作最基本的一个原则：不仅仅是要实验结果好，更重要的是搞明白为什么会好。亦师亦友，这个词再合适不过了。

感谢环绕智能实验室的兄弟姐妹们，虽然大部分都是实习学生，但是我们之间并没有任何距离。高卫强、朱云、高永强、李伟浩，很珍惜与你们周六的下午一起踢球的时光；孙剑同学，你的口头禅已经不止一次地成为了实验室的流行语，很看好你，将来一定能成大器；宋阳、李阳、戚恩，三位师弟，我从你们身上也学到了很多；还有王育红，感谢你为我标注的数据，没有你的工作，不可能有这些研究成果；还有实验的曾经的和现任的同学们，感谢你们为艾尔斯口语考试系统做出的努力，感谢你们在学习和生活中对我的帮助和支持，感谢你们陪我一起“二”的那些时光，有你们这样的朋友，是我此生的幸运！

那些年改过的论文

那些年熬过的深夜

那些年，我们一起搞过的科研……

摘 要

英语发音自动评测技术作为计算机辅助语言学习（Computer Assited Language Learning）系统在语音评测领域的一个重要应用，有着积极的研究意义和应用前景。本文针对英语口语考试中的朗读题型和复述题型，研究了文本相关和文本无关的英语发音自动评测技术。并基于研究成果开发了艾尔斯口语考试系统。

对于朗读题型的自动评测，区别于传统基于强制切分技术的评分流程，本文提出了基于强制对齐、标注语音段、非监督性说话人自适应、语音识别以及单音素循环网络解码的评分流程，在此基础上提取了反映发音质量、流利度和内容相关的评分特征，最后利用多元线性回归拟合方法，得到机器评分结果。实验结果表明，本文的提出的方法获得了 0.923 的机器评分与人工评分的相关度，超过了传统基于强制对齐技术的朗读题型自动评分方法。

对于复述题型的自动评测这一前沿课题，本文也做出了积极探索，将研究重点放在内容相关（content related）评分特征的提取及复述风格语言模型的构建两个问题。对于前者，本文提出了基于 Similarity、Keywords Coverage Rate (KCR)、Word Number (WN) 和 Unique Word Number (UWN) 四维内容相关的复述题型评分特征集，结合传统的反映发音质量和流利度的评分特征集，获得了 0.621 的机器评分与人工评分相关度，超过了传统评分特征集，证明了本文提出的内容相关的评分特征集的有效性。对于复述风格的语言模型的构建问题，本文选择和收集了话题相关文本语料、口语风格文本语料和书面风格文本语料，采用线性插值方法，提出了基于混合语言模型的语言模型自适应方法。实验结果表明，相比传统的语言模型建模方法，本文提出的方法最高将困惑度降低了 61.6%，并将复述题的语音识别的单词错误率降低了 20.7%，充分证明了该方法的有效性。

最后，基于朗读题型和复述题型的自动评测技术，本文开发了艾尔斯口语考试系统，并将系统推广至中学生日常英语教学中。

关键词：计算机辅助语言学习 自动评分 朗读题 复述题 语言模型自适应

Research on Automatic Scoring of Pronunciation in Oral English Test

Meng Chen (Computer Application Technology)

Directed By Helen Meng, Lan Wang

As an important application of Computer Assisted Language Learning (CALL) in the field of oral proficiency evalucation, there are great significances for automatic scoring of English pronunciation both on research and application. Based on the tasks of reading aloud and retelling in oral English test, we researched automatic scoring technology for text-dependent and text-independent tasks. And we developed the Interactive English Learning System (IELS) based on the research.

For the task of reading aloud, we proposed a novel scoring flow by process of forced-alignment, segmentation, unsupervised speaker adaptation, speech recognition, and phoneme loop recognition, which is different from the traditional scoring flow. Then we extracted the scoring features which represent intelligibility, fluency and content realted. Linear models are introduced to combine different features for automatic scoring. Experimental results show a correlation of 0.923 between machine scores and human scores can be obtained, which is higher than the traditional method.

For the task of retelling, we focused on extracting content related scoring features and language modeling for retelling speech. For the first problem, we proposed a novel set of content related scoring features which consist of Similarity, Keywords Coverage Rate (KCR), Word Number (WN) and Unique Word Number (UWN). By combining the traditional scoring features of intelligibility and fluency, we obtained a correlation of 0.621 between machine scores and human scores, which is higher than traditional scoring features. For the second problem, we proposed a novel language modeling mothod using mixture models. We conducted the language modeling by collecting topic related, spoken-style, and document style related text sources firstly, then adapting the language model using mixture models by interpolation. Experimenal results show that up to 61.6% relative reduction of perplexity and 20.7% absolute reduction of word error rate have been obtained by our best performing model.

Furthermore, we developed the Interactive English Learning System (IELS) based on the automatic scoring technology. And we also deployed the system in high school, which had a positive significance.

Keywords: CALL, Automatic Scoring, Reading Aloud, Retelling, Language Model Adaptation

目 录

图目录

表目录

第一章　绪论

1.1 研究背景

当今社会，随着国际化交流的日益频繁，人们对外语学习的需求日益增加，掌握一门外语已经成为现代社会的一项重要技能。在中国，英语学习已经成为了教育领域的一个重要的组成部分。英语课程已经成为了从小学到研究生阶段的必修课程。而在教育领域之外，出国留学、商务英语也包含着大量的英语学习者，这些都印证了社会对外语学习的巨大需求。

在外语学习中，口语学习的重要性更是不言而喻。口语交流作为最为直接的交流方式，相比于读写能力，更为必须。与此同时，口语学习也是外语学习中最困难的部分。口语学习需要交互式训练，需要有效的反馈，这些都构成了口语学习者的巨大障碍。在中国，英语口语学习的困境显得更为突出，尽管广大英语学习者历经了从小学到高等教育将近十几年的英语课程学习，却仍然摆脱不了“哑巴英语”的结局，这说明，传统的英语口语教育存在着一些弊端。首先，由于英语师资力量的欠缺，传统英语口语学习很难做到一对一的教学辅导。而英语口语学习需要专业的反馈，包括发音质量、语速、词汇、语法等方面的有效反馈，这显然是传统英语口语教学无法保证的。其次，由于传统英语口语教学的时间局限性，口语学习者缺乏充足的练习时间。而单独的练习又显得枯燥、乏味，很难激起口语学习者的热情。这些都是口语学习者巨大的需求与传统口语教学方式之间的矛盾。

此外，传统口语考试也存在着不少弊端。首先，传统的口语考试耗时耗力。组织一次中考或者高考英语口语考试，往往需要动用全省的英语教师资源，当学生规模较大的情况时，考官更是疲于应付，很难做到有效评分。其次，传统的口语考试不够公正。由于大规模的口语考试时间跨度比较大，先考完的考生往往容易透露考题，这对于后面的考生显然具有一定的优势，从而考试缺乏公正性。最后，传统的口语考试评判结果波动较大。考官在评判过程中，一方面有可能会受到前面的考生的外语水平以及当前考生的举止、外貌等因素的影响，另一方面，由于考官自身的情绪波动，对于同样的考生，在不同的时期，考官也可能会给出不同的成绩。这些，都在呼吁着一种更为公平、公正的英语口语考试评判方式的出现。

计算机辅助语言学习（Computer Assisted Language Learning，CALL）是指运用计算机替代（或辅助）人进行传统的语言教学任务，达到一对一教学、个性化教学的目的。发音自动评测技术则是计算机辅助语言学习在语音评测中的重要应用。在口语学习方面，它能辅助或者替代教师，对学生的口语发音进行评价，并将信息反馈给学生，帮助其纠正错误发音，从而大大提高口语学习效率。在考试方面，它依赖于机考系统，能辅助或

替代老师进行更加客观、公正、高效的评分。

与传统的口语考试相比，发音自动评测系统展现出不可替代的优势：首先，它实现了考官与考生的分离，避免了考官因和考生的关系而造成不公正的给分；其次，它能自动保存语音，以方便审查，特别对于口语考试而言，避免了传统考试的不可逆性；另外，它支持所有考生使用同样的试题、同时考试、同时录音，这样不仅严格公平，并且方便了大规模组织，还避免了传统的排队式口语考试情况下，先考完的考生对未考的考生泄题的情况。最后，将智能化的口语评分融入机考系统，不仅仅能节省大量的人力物力资源，还能大大提高口语评分的评分效率和客观性。此外，基于发音自动评测的英语口语学习系统还能提供更加丰富和个性化的反馈信息，这些都对英语口语学习者具有巨大帮助。

总而言之，将计算机辅助语言学习技术应用到口语学习中，尤其是口语评测中，可以有效地解决当前口语考试的诸多问题，对广大口语学习者具有极大的帮助。

1.2 研究内容

发音自动评测按照评测题型可分为文本相关和文本无关两种形式。

- **文本相关题型：**要求考生的回答内容与标准答案完全一致，典型题型包括朗读题和跟读题。对于朗读题型而言，测试者需要按照给定的文本提示，在给定时间内朗读出来。而评测的标准则是基于朗读的完整度、流利度等指标进行衡量。主要侧重于考察考生的发音、语调和韵律等基本语言能力。
- **文本无关题型：**该类题型不要求考生的回答与标准答案完全一致，只需主题相同或者表达的意思相同即可。典型题型包括复述题和口头作文。对于复述题型而言，测试者先听一段小故事，随后用自己的话进行复述，要求涵盖到小故事的要点。这类题型更注重对考生口语能力的综合考察，包括语音、语法、词汇、语调以及韵律等综合语言能力。

本文研究的主要内容是关于文本相关题型和文本无关题型的自动评测技术。具体包含了朗读题的发音自动评测技术和复述题的发音自动评测技术的研究。其中，两类题型的介绍如下：

朗读题的主要形式是：播放一段带有字幕的英语视频（含字幕，100~150 词），然后要求考生跟着视频的播放顺序，按照字幕的提示进行朗读。朗读时长 1 分钟。

复述题的主要形式是：播放一段英文音频（无字幕，200~300 词），音频内容通常是描述一个小故事，然后要求考生根据听到的内容进行复述，复述可以与音频内容完全一样，也可以按自己的方式表达，但是基本内容要求与音频一样。回答时长 2 分钟。

下面将介绍国内外在发音自动评测领域的研究现状。

1.3 国内外研究现状

对于发音评分的研究国内外已经开展了二十多年的研究。为了提供不同的反馈信息，需要开展不同粒度层次的发音质量评分。目前这些粒度层次包括音素级别、句子级别以及说话人级别。普遍的方法都是基于语音识别技术。描述发音质量的特征一般从语音识别器的输出中提取，特别是强制对齐（Forced Alignment）和识别结果（Recognition）。例如，后验概率、音素模型置信度、时长、语速、识别率以及停顿时长等。为了提取这些发音特征，通常需要准备参考文本、声学模型，目标语言的音素时长统计知识，以及语言模型等。

发音评分可分为文本相关（text-dependent）和文本无关（text-independent）两种形式。文本相关的发音评分需要发音人按照给定文本进行发音，同时，给定文本也是评测发音人发音水平的标准；例如朗读、跟读的评分。文本无关的语音评测中，文本不是必要的，即使有文本，也仅仅是一种参考，发音人需要用自己的语言表示题目所要求的内容。如翻译题、复述题、口头作文题。

近年来，文本相关的发音评分是研究主流，其中比较典型的工作是 SRI International（Speech Technology and Research Laboratory）的 Franco 等人从事的工作 [1] [2] [3] [4]。他们的系统可以从音素级别和句子级别评估母语和非母语说话人朗读英语的发音质量。说话人按照给定文本进行朗读，通过强制对齐技术，将语音信号与 HMM 理想的解码路径对齐。基于强制对齐的结果，提取反映发音质量的评分特征。他们发现，音素的后验概率（posterior probability）、时长得分（duration score）、音节时长（syllabic timing）等特征与人工评分存在着较高的相关度。他们还对比了不同的评分方法，包括线性回归、非线性回归（人工神经网络、回归树模型）等，将不同特征进行组合，得到最终的机器评分。

Cucchiarini 等人沿着与 Franco 等人类似的方法开发了荷兰语发音自动评分系统 [5]。但是，他们采用了更多的评分特征，包括 HMM 置信得分、各种时长得分、以及语言中的暂停、词重音、音节结构以及语调等信息。最终，机器评分与人工评分的相关度达到了 0.67~0.92。

Bernstein 则为 SET-10 英语口语考试开发了一套自动评分系统 [6] [7]。其中，题型包括：朗读、句子跟读、造句、反义词、简短问答以及开放性问答。除了最后一类题型采用人工评分，其他题型均利用机器自动评分。他们开展了关于 SET-10 英语口语考试通用性的实验，最终得到的结论是：SET-10 考试分数可以反映欧洲口语交互委员会（Oral Interaction Scale of the Council of Europe）中描述的非母语（第二语言）说话人的英语口语水平。这篇文章还指出，他们的系统的机器打分与人工评分的相关度达到了 0.73~0.88。

对文本无关的发音评分目前开展的并不多，尤其是对非母语说话人的低限制度（unrestricted）、自发性发音（spontaneous speech）的自动评分更少。在这方面，Zechner 和 Bejar 进行了一些探索性工作 [8] [9]。他们主要针对的是 TOEFL 考试中简短问答题型

（回答在 1 分钟左右）的自动评分。他们的主要工作集中在新的流利度评分特征的提取，在评分方法上，他们对比了分类回归树（CART）和支持向量机（SVM），结果发现 SVM 在定量分析方面具有优势，而 CART 则对于数据中潜在的模式具有更透明的展示。

最近， ETS（Educational Testing Service）针对托福口语考试中自发性描述题型，开展了更为细致和实用的研究 [10] [11]。他们的系统利用非母语英语语料库训练语言识别器，通过评分特征计算模块，从识别结果中计算了一系列基于流利度的评分特征，然后通过多重线性回归的方法预测发音评分，最终在 TPO（TOEFL Practice Online）语料上取得了 0.57 的机器分与人工评分相关度。虽然这与人工之间评分相关度（0.74）相比还有一定的差距，但是对于低利害关系的训练环境而言，他们的系统已经达到了实用水准，因为系统的自动评分标准覆盖了诸如流利度、词汇、语法以及发音质量等多方面的体现交流能力的特征。

在国内，科大讯飞也做了大量的工作 [12] [13]。严可在英文朗读题和复述题的自动评测中做了一些技术研究。朗读题评测方面与前人的工作基本类似，而复述题评测方面主要创新之处在于根据范文定制语言模型，进行语音识别。同时引入基于词图的自动评分特征，采用了诸如关键词覆盖率、范文单词覆盖率和用词变化程度等基于内容的评分特征，取得了不错的评分性能。

此外，中国科学院自动化所的江杰、徐波等人在口语测试的自动评估技术做的大量工作也值得借鉴 [14] [15] [16]。江杰以汉语普通话水平测试作为研究对象，重点研究了朗读题型的评估技术、发音错误自动诊断技术和问答题型的评估技术。尤其是在问答题型的自动评估方面，提出了基于语义的自动评估技术，为文本无关的语音评测做出了较为深入的探索。

在系统方面，上述代表工作都有典型的系统，如 Franco 等人的 VILTS 系统在说话人级别的发音评分取得了不错的效果； CMU 大学的 LISTEN 项目[17]在帮助小孩学发音上取得了良好的效果；ETS 的 SpeechRater 系统已经在托福口语考试中达到了实用水平；国内，科大讯飞公司也推出了畅言互动英语校园学习平台 [18]，将英语学习、口语考试、发音自动评分融合成一套实用性很强的学习系统；自动化所将也将发音自动评测技术应用于国家普通话水平测试以及中考英语口语考试中。

1.4 本文的主要工作和创新点

本文主要探讨朗读题型和复述题型的自动评测问题。首先，对文本相关的朗读题型自动评测技术进行研究，随后，探索目前比较困难的文本无关的复述题型自动评测技术，并基于研究结果开发了艾尔斯口语考试系统。

本文首先着手于朗读题的自动评测技术的研究。仔细分析了朗读题型的特点以及人工评分的考察要点，提出了一个快速有效的自动评分机制，包括强制对齐、标注语音段、非监督性说话人自适应、语音识别以及单音素循环网络解码。选取了反映发音质量、流

利度和内容相关的评分特征，包括 GOP 得分，EPN（音素错误数量），ROS（语速），SPR（停顿时长比例），Corr（正确率）和 Acc（准确率），运用线性拟合的方法对人工评分进行学习，并预测考生得分，在 READ-DB 数据库上取得了与人工评分相关度为 0.923 的评分结果。

然后，本文探索了复述题型的自动评测技术，提出了内容相关评分特征，包括 Similarity（内容相似度）、KCR（关键词覆盖率）、WN（单词数）和 UWN（非重复单词数），结合传统反映发音质量和流利度的评分特征，运用线性拟合的方法对人工评分进行学习，并预测考生得分，在 RETELL-DB 数据库上取得了与人工评分相关度为 0.621 的评分结果。随后，本文将研究重点转移到提高复述题型中非母语说话人自发性表述的语音识别率，提出了基于混合语言模型的语言模型自适应方法，并基于该方法构建了复述风格的语言模型，上述方法应用于 XIXIANG 和 YUCAI 语音数据集，语音识别的单词错误率分别降低了 16.9%和 20.7%。

最后，基于英语发音自动评测技术的研究成果，本文还开发了艾尔斯口语考试系统，并将系统推广至中学生日常英语教学中。

本文的创新点包括：

- 对于朗读题型自动评测问题，本文提出了基于强制对齐、标注语音段、非监督性说话人自适应、语音识别以及单音素循环网络解码的评分流程，取得了优于传统评分流程的评分效果；
- 对于复述题型自动评测问题，提出了包括 Similarity（内容相似度）、KCR（关键词覆盖率）、WN（单词数）和 UWN（非重复单词数）4 维内容相关的复述题型评分特征，并融合传统基于发音质量和流利度的评分特征，取得了优于传统评分特征的评分效果；
- 对于复述题型自动评测中提高语音识别率的难点，本文提出了基于混合语言模型的语言模型自适应方法，并将方法应用于复述风格的语言模型构建中。实验证明，相比于传统的语言模型建模方法，该方法构建的语言模型的困惑度得到了大幅度降低，语音识别的词错误率也得到了大幅度降低。

1.5 论文的组织结构

本文在朗读题型自动评测和复述题型自动评测方面做了大量的工作，这里将简单陈述本文的组织结构。

第二章为发音自动评测技术综述。包括语音识别技术原理简介、发音自动评测技术介绍以及非母语说话人英语口语表述的典型特点。语音识别技术主要介绍了隐马尔科夫模型的基本概念、语言模型、说话人自适应技术以及主流的语言模型自适应方法；发音自动评测技术介绍了发音自动评测的基本框架及主要组成部分。最后，从语言学的角度介绍了非母语说话人（特别是中国人）英语口语表述的特点。

第三章详细介绍了朗读题型的发音自动评测的实现。首先分析了朗读题型的特点和难点，其次分析了朗读题型自动评测中用到的语音数据库，随后详细阐述了朗读题型自动评测流程的搭建，最后是实验结果和分析以及改进方案。

第四章详细介绍了复述题型的发音自动评测的实现。首先分析了复述题型的特点和难点，其次分析了复述题型自动评测中用到的语音数据库，随后详细介绍了复述题型自动评测流程的搭建，并着重介绍了内容相关评分特征以及复述风格语言模型构造方法，最后给出了实验结果。

第五章中，我们将给出艾尔斯口语考试系统的简要介绍以及系统截图，随后是全文的总结，最后是对未来工作的展望。

第二章 发音自动评测技术综述

2.1 语音识别技术简介

现有的 ASR（Automatic Speech Recognition）系统都建立在统计模式识别的基础之上。假设我们用随机向量 O 来表示一串未知的语音信号，定义为：

$$O = o_1, o_2, \dots, o_T \tag{2-1}$$

其中o_T是在时刻 T 的语音观察序列。用向量$W = w_1, w_2, \dots, w_M$代表语音信号所对应的文本内容，则 ASR 系统的任务便是在给定声学信号 O 的前提下找到最有可能的文字序列 W。即求$arg\max P(W|O)$利用贝叶斯公式可以得到：

$$\widehat{W} = arg\max_W P(W|O) = arg\max_W \frac{P(W)P(O|W)}{P(O)} \tag{2-2}$$

显然，为了找到最有可能的词序列$\widehat{W}$，必须要先计算P(W)和P(O|W)。其中第一项代表了独立于观察信号的文本的先验概率，称之为语言模型。第二项代表给定文本的前提下，观察序列出现的概率，称之为声学模型。

2.1.1 HMM 简介

我们知道，随机过程又称随机函数，是随时间而随机变化的过程。在马尔科夫模型中，每个状态代表了一个可观察的事件，所以，马尔科夫模型有时又称作可视马尔科夫模型（Visible Markov Model，VMM），这在某种程度上限制了模型的适应性。在隐马尔科夫模型（Hidden Markov Model，HMM）中，我们不知道模型所经过的状态序列，只知道状态的概率函数，也就是说，观察到的事件是状态的随机函数，因此，该模型是一个双重的随机过程。其中，模型的状态转换过程是不可观察的，即隐蔽的，可观察事件的随机过程是隐蔽的状态转换过程的随机函数。

语音是一个随机过程，每次发音时，我们都可以得到一个帧矢量序列（称为观察序列）

$$O = \{o_1, o_2, \dots, o_T\} \tag{2-3}$$

对于同一词的不同发音，O 的帧数 T 和o_i都在变化，可以看成是该随机过程模型的多次实现。从语音产生的过程来看，可以想象为声道沿不同位置转移时，每一位置产生一个随机声学输出。可以把各声道位置想象成各个状态q_i，而观察序列可想象为在该状态下的一个随机输出o_i。因此，语音的随机过程可以看成是两个随机过程构成的：

一是状态转移的随机过程；二是输出的随机过程。

可见，隐马尔科夫模型很好的描述了语音随机过程。

HMM 中有三个基本问题：

- **评价问题：**给定一个观察序列$O = o_1 o_2 \ldots o_T$和模型μ，如何快速地计算出给定模型μ情况下，观察序列 O 的概率，即P$(O|\mu)$？
- **解码问题：**给定一个观察序列$O = o_1 o_2 \ldots o_T$和模型μ，如何快速有效地选择在一定意义下“最优”的状态序列$Q = q_1 q_2 \ldots q_T$，使得该状态序列“最好地解释”观察序列？
- **学习问题：**给定一个观察序列$O = o_1 o_2 \ldots o_T$，如何根据最大似然估计来求模型的参数值？及如何调节模型μ的参数，使得P$(O|\mu)$最大？

经过多年的研究，前人已经针对这三个问题分别给出了较好的解决方案。目前的主流方法是：使用前向算法进行评价问题的计算，使用 Viterbi 算法解决解码问题，使用 Baum-Welch 算法来解决学习问题。

2.1.2 语言模型

一个语言模型通常构建为字符串 W 的概率分布 P(W)，这里 P(W)试图反映的是字符串 W 作为一个句子出现的频率。对于一个由m个基元(“基元”可以为字、词或短语等)构成的句子W $= w_1 w_2 \ldots w_m$，其概率计算公式可以表示为：

$$P(W) = \mathrm{P}(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \ldots, w_{i-1}) \tag{2-4}$$

在公式(2-4)中，产生第$i(1 \le i \le m)$个词的概率是由已经产生的$i-1$个词$w_1 w_2 \ldots w_{i-1}$决定的。一般地，我们把前$i-1$个词$w_1 w_2 \ldots w_{i-1}$称为第i个词的“历史（history）”。这样，随着历史长度的增加，不同的历史数目呈指数级增长。为了解决这个问题，可以将公式(2-4)简化为(2-5)：

$$P(w_1, w_2, \ldots, w_m) \backsimeq \prod_{i=1}^{m} P(w_i | w_{i-n+1}, \ldots, w_{i-1}) \tag{2-5}$$

其中 n 的值通常为$1 \le n \le 4$，这样的模型我们称之为 n 元文法（n-gram）。

N 元文法的估计通常建立在最大似然估计的基础上，通过计算训练文本中某个词出现的次数，我们可以得到：

$$P(w_i | w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})} \tag{2-6}$$

其中，$\mathrm{C}(\cdot)$就是训练文本中某个词或词序列出现的次数。由于训练文本的局限性，上述公式的分子可能为 0，这种情况下，$P(W)$也会为 0。在语音识别中，这意味着无论声学信号是什么，结果都为 0，显然这是不合理的。

因此我们需要平滑算法，通常的平滑模型包括后备模型（back-off model）和插值模型（interpolated model）。

评价一个语言模型最常用的度量就是根据模型计算出的测试数据的概率，困惑度

（perplexity）常常用来衡量语言模型的优劣。困惑度 Perplexity 的计算如下：

$$Perplexity = P(w_1, w_2, \dots, w_m)^{-\frac{1}{m}} \tag{2-7}$$

通常，困惑度越小越好，在英文文本中，n 元文法模型计算的困惑度范围大约为 50 至 1000 之间，具体指与文本的类型有关 [19]。

2.1.3 说话人自适应技术

说话人自适应是为了弥补训练和测试发音风格、音色等差异，另外，自适应本身也可以弥补部分环境差异。因此，说话人自适应是非常有用的加强语音识别系统鲁棒性的方法。说话人自适应分为监督性自适应和非监督性自适应，前者是指已知正确的自适应数据对应的脚本信息，后者是指未知自适应数据对应的脚本信息。

目前自适应方法主要采用最大似然线性回归的方式（Maximum Likelihood Linear Regression，MLLR），因为它能在数据量较少时取得较好的性能。MLLR 方法是根据最大似然准则，通过一系列的线性变换来调整模型。HMM 可由均值和方差来描述，对方差做 MLLR 变换计算复杂且性能提升有限，另外由于自适应训练数据往往较少，所以一般仅对各 HMM 作均值变换。

在 HTK [20]中，非监督性自适应通过以下三个步骤完成：

- 通过 HVite 做第一次 Viterbi 解码，得到识别结果，即音频对应的脚本信息；
- 通过 HERest，利用第一次解码的结果，估计转移矩阵 Transforms；
- 利用第二步得到的转移矩阵，进行第二次 Viterbi 解码，得到最终的识别结果。

2.1.4 语言模型自适应技术

在语音识别系统中，语言模型的性能好坏直接影响了整个系统的性能。尽管语言模型的理论基础已经比较完善，但在实际应用中由于自然语言的多变性，常常会碰到一些难以处理的问题。这些多变性包括：

- 语言在进化过程中常常出现新词汇。
- 不同领域的语言中词序列统计分布往往显著不同。
- 人们总是习惯于根据手头的任务调整语言的使用。比如，人们用在正式的科技论文和日常的电子邮件中的语言的句法结构显然不同。
- 受社会经济地位、情感音素的影响，人们的论述风格各异，尤其是对于口语化的自然语言更为明显。

由于自然语言内在的多变性，训练语料和识别任务中的语言在词法、句法和语义特征上很有可能不同，从而造成了语言模型对跨领域的脆弱性。这会严重地降低语音识别的识别率。因此，为了提高语言模型对语料的领域、主题、类型等因素的适应性，R. Kuhn 等人和 J. Kupiec 提出了自适应语言模型的概念。在随后的这些年里，人们相继提出了一系列的语言模型自适应方法，并进行了大量的实践。

各种各样的语言模型自适应方法可以大体分为三类：（1）插值模型，例如动态缓冲模型；（2）约束规范模型，例如基于最大熵的语言模型；（3）元信息提取模型，这里的元信息指的是语料词序列中观察不到的隐含的知识信息，例如潜在的表述话题、语义和句法信息等等。比较典型的方法是基于隐含语义分析的语言模型自适应方法[21]。
尽管各种自适应方法各异，但是通用的语言模型自适应框架都可以用图 2.1 描述：

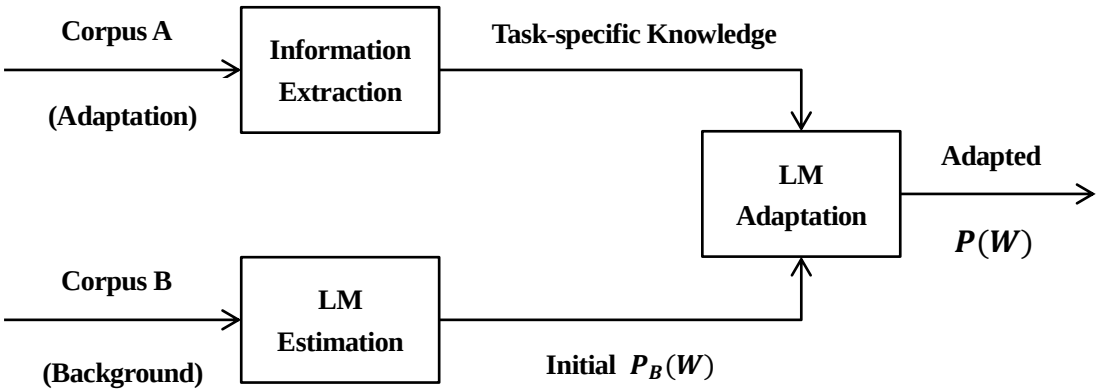

图 2.1 统计语言模型自适应方法的通用框架

其中 Corpus A 指的是与当前识别任务相关的小型的自适应语料库；Corpus B 则是一个大型的背景语料库，有可能包含了不同任务的语料。Corpus B 用来训练一个最初的概率估计$P_B(W)$，尽管这个背景统计语言模型的风格与实际的识别任务可能不匹配；Corpus A 则用来提取一些与当前任务相关的特定的信息，例如缓存中的常用词频率、约束信息以及话题实体等等。自适应的思想便是如何基于 Corpus A 中提取的信息来动态的修改背景统计语言模型中的概率估计。

在本文中，由于复述题自动评测中面对的是非母语说话人自发性表述，因此在这里更加关注非母语说话人自发性表述的语言模型自适应方法。

非母语说话人自发性表述在语言模型建模上有如下特点：

- 自发性表述与正常规范文本语料在表达方式上不匹配；
- 可信的自发性表述转写文本语料不够充足；
- 表达方式高度依赖说话人。

针对这些特点，语言模型的改进方法可分为以下四类：

- 合成模型：如果能获得大量自发性表述风格的文本语料，可以将这部分语料与话题相关或领域相关的书面风格文本语料进行组合，训练语言模型 [22] [23] [24]。
- 词类模型：如果文本语料并不是很充足，基于词类的语言模型可以很好的满足话题相关的语言模型构建。其基本思想是将语料中出现的词进行聚类，然后构建基于词类的 N-gram 语言模型 [25] [26]。

基于词类的语言模型是对基于词的语言模型的改进。假设词属于类，由语言模型计算 n 元概率$\mathrm{P}(w_i|w_{i-n+1}^{i-1})$，公式如下：

$$\mathrm{P}(w_i|w_{i-n+1}^{i-1}) = \mathrm{P}(w_i|C_j)\mathrm{P}(C_j|C_{j-n+1}^{j-1}) \tag{2-8}$$

一般情况下，由于词类的数目小于词的数目，这样，在估计 n 元概率时面临的数据稀疏问题在一定程度得到缓解，提高了对训练语料中未出现的词串的预测能力。此外，还压缩了语言模型的尺寸。

如果把每个词看成一个类就回退到基于词的语言模型了，因此基于词的语言模型可以看成是基于词类的语言模型的一个特例。

- 转换模型：如果同时有自发性表述风格的文本语料以及对应的书面风格文本语料，可以基于平行语料库估计书面风格文本与自发性表述风格文本之间的转移概率，从而借助机器翻译的概念，通过书面风格文本语料自动生成自发性表述文本语料。使用这种方法的前提是需要具备书面风格文本和自发性表述文本的平行语料，并且进行词法分析[27]。
- 填充词预测模型：针对自发性表述中包含的大量不流利现象（例如填充词），基于填充词的位置进行建模，预测填充词可能出现的位置，通过迭代转写的方法，由规范文本语料预测生成包含填充词的自发性表述语料[28] [29]。

2.2 发音自动评测技术简介

发音自动评分基于自动语音识别（Automatic Speech Recognition, ASR）技术，通过提取反映学习者口语表述能力的评分特征，结合专业英语老师的人工评分，借助机器学习的方法，训练评分模型，基于评分模型预测考生发音得分。其基本框架如图 2.2 所示：

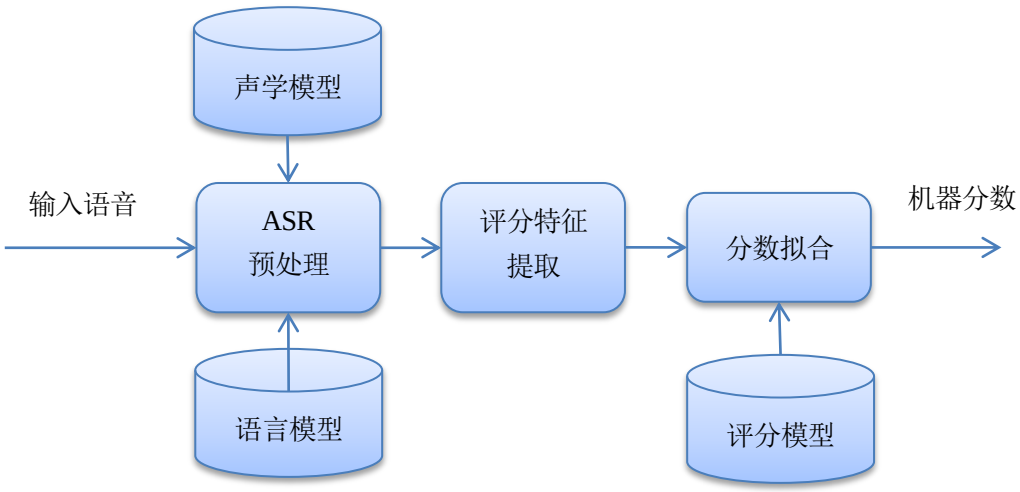

图 2. 2 发音自动评分框架

- ASR 预处理：利用语音识别器，对语音进行的一些操作，目的是获取尽可能多的反映考生口语水平的信息。文本相关的发音自动评测通常进行强制对齐（Forced Alignment），获取测试语音相对于朗读文本的对齐信息，匹配语音段的时间和置信

度。文本无关的发音自动评测通常进行语音识别（Recognition），获取测试语音的内容及时间信息。

- 评分特征提取：从识别器的输出信息中提取表征测试者口语水平的特征。目前常用的特征可以分为发音质量特征、流利度特征和内容相关特征。发音质量特征一般是通过在自动语音识别过程中加入置信度计算而得到的，它表示测试语音相对于给定的模型的不确定程度。流利度特征一般指与时间相关的特征，比如语速、停顿时长等。内容相关特征一般指识别结果与可能的参考答案之间相似度，比如内容正确率、关键词覆盖率等。
- 分数拟合：通常指的是训练评分模型以及预测分数的机器学习算法。比较常见的有线性回归模型、支持向量机和人工神经网络等。
- 声学模型：用母语说话人或者非母语说话人语料库训练而成的隐马尔科夫模型。用于语音识别。
- 语言模型：用与任务匹配的文本语料训练而成包含词条统计信息的 N 元文法统计模型，用于语音识别。
- 评分模型：对含有人工评分标注的测试语音提取评分特征，结合评分特征和人工标注信息，利用分数拟合阶段的机器学习算法训练的统计模型。用于预测机器评分。

2.3 非母语说话人英语口语表述的典型特点

本节将从语言学的角度介绍非母语说话人（主要是中国人）在英语口语表述上的典型特点[30] [31]。

非母语说话人在英语口语表述中存在着大量的非流利现象。根据《语言与语言学词典》，非流利指任何形式的言语流利故障（包括病理性的言语产出失调）。杨军认为，非流利指的是在时定、韵律和语序等方面明显区别于流利话语的口语产出。并将非流利类型分为停顿、重复、自我修正三类。戴朝晖利用 PACCEL-S 语料库，研究中国英语专业大学生汉英口译中的非流利现象，将非流利类型的三种典型类型做了更为细致的划分。

停顿是指语流中断现象，包括无声停顿（unfilled pause）和有声停顿（filled pause）。前者是指语流中的无声或沉默，后者指说话者“由于无法或不愿意产出所需的词但证明说话者仍处语言活动中的有声证据”。比如，um，uh，er 和 I mean，well，you know 等。

非流利性重复是指单词、词组或句子被重说一遍，且不对句法、词形或句序做任何改动，它可以是完整的音、音节、词段、词组、词串甚至句的重复。例如，在 give him give him good food 中，give him 说了两遍。重复频数计算不包括为了修辞效果而所作的重复，如 very very angry，也不包括语篇上下指称性重复。

自我修正可分为三类：不同信息修正（Different Repairs）、恰当修正（Appropriacy Repairs）和错误修正（Error Repairs）。不同信息修正指的是用不同的信息替换当前信息；恰当修正指的是用恰当的表达方式替代当前表达方式以消除歧义，使表达更加精确或保

持内容的前后连贯，或使语用恰当正确；第三类是错误修正，即纠正当前表达方式中的词汇、句法和语音方面的错误。下面的三句话分别是三类自我修正的示例：

You have to we have to make a contract.（说话者意识到用 you 不合适，实际上是餐饮签订合同，故转用 we）

There are er er twenty er tables er about twenty tables.（说话者意识到给出的数据不准确，因此加入了 about 一次）

It was nice to meet you and that you choose you chose us.（说话者意识到时代不正确，转用 chose）

2.4 小结

本章系统地介绍了发音自动评测技术所涉及到的语音识别技术和发音自动评测技术，并从语言学的角度分析了非母语说话人（主要是中国人）英语口语表述的典型特点。其中语音识别技术介绍了隐马尔科夫模型、语言模型的相关概念、说话人自适应技术，并重点介绍了语言模型自适应技术。发音自动评测技术介绍了发音自动评分的基本框架以及框架中的各个组成部分。非母语说话人英语口语表述的典型特点则重点介绍了停顿、重复和自我修正三种常见的不流利现象。这一章涉及到的理论知识是后面两章搭建自动评分系统的基础。

第三章　朗读题发音自动评测技术

本章将详细介绍朗读题型的发音自动评测技术。我们将基于对朗读题型的特点、考生发音特点的详细分析，给出朗读题型自动评分流程的搭建过程，并在数据库上给出实验结果和改进措施。

3.1 朗读题型特点及难点分析

3.1.1 朗读题型简介

朗读题型是口语测试中的常见题型。在各种水平、不同类型的口语测试中，几乎都可以看到朗读题型。对于初级的口语水平测试，朗读题型主要用于评估测试者的发音、语速、声调、重音等基本口语能力，因此，朗读内容的选取一般都会偏向于简短易懂的文章片段。在较为高级的测试中，朗读题型则会更加看重韵律等高层次语言能力的掌握，因而朗读内容的选取会更加复杂化。目前对于朗读题型的评估技术，主要还是集中在初级的口语水平测试上。

本文研究的是广东省高考英语口语考试的朗读题型，其形式如下：系统播放一段带有英文字幕的英语视频，时长为两分钟，100~150 词，随后要求考生跟着视频的播放顺序，按照字幕的提示进行朗读。朗读录音时长 1 分钟。为了全面考察考生的发音，朗读题会包含部分生词、缩写、数字以及专有名词等。图 3.1 是朗读题的示例：

PART A Reading Aloud

And on the northern boundary, the ice-capped peaks of the Himalayas. This mountain ranges the highest and fastest-growing on the planet. It is also the source of a river that lies at the heart of the Indian nation—the Ganges. Over 1,500 miles in length, the Ganges flows out across the vast plains of northern India. It supports the lives of 1/12 of the world's population. But for India's 800 million Hindus, the Ganges is more than just a river. She is a goddess known as Ganga. Her waters have the power to purify the soul and cleanse the spirit. To bathe in the Ganges is one of the most profound spiritual experiences a Hindu can have. （117 words）

图 3. 1 朗读题型示例

3.1.2 朗读题型考察要点

朗读题型同时给定了朗读文本和示范读音，主要侧重于考生的发音质量。因此内容的准确性是朗读题型考察的重点，包括少量生词以及缩写、数字及专有名词的发音考察。此外，由于朗读题型的文本是篇章级别的，并且要求考生按照视频的字幕提示的节奏进

行朗读，因此，语速也是考察要点之一。考生每次朗读的单元是句子级别，重读和语调也属于考察的部分。对于初级的口语水平测试，朗读题型主要用于评估测试者的发音、语速、声调、重音等基本口语能力；而对于较为高级的测试，朗读题型则会更加看重韵律等高层次语言能力的掌握，因而朗读内容的选取会更加复杂化。

3.1.3 朗读题型评分标准

基于上述的朗读题型考察要点的分析，朗读题评分标准需要综合考虑考生答案的流利度（Fluency）、连贯（Coherence）、发音（Pronunciation）、语速（Speech Rate）、准确度（Accuracy），基于各项内容加权表征考生的口语水平。总分为 20 分，分为 ABCD 四个等级。具体评分细则如表 3.1 所示：

表 3.1 朗读题型评分标准

Feature / Grade	流利度，连贯 (Fluency&Coherence)	发音 (Pronunciation)	语速 (Speech rate)	准确度 (Accuracy)
A (17-20)	表达流畅，连贯，自然；逻辑清晰	重音和音韵准确；利用不同语调表达意义；单独词汇发音准确,容易被理解	按照原语速，并掌握节奏和停顿	几乎没有词汇、发音、句法或语言错误
B (13-16)	表达基本流畅，连贯，自然；逻辑清晰，但有个别处影响自然语流	发音准确，利用重音和语调，但有一些语法错误	基本按照原语速，并掌握一定的节奏和停顿；至多疏漏 3 个单词	只有极少量的词汇、发音、句法或语言错误
C (9-12)	尽管有些较长的停顿，但能够有节奏地连贯表达较长的句子	母语音较重，导致一些重音和语调使用不当；但不影响理解	基本按照原语速，并掌握一定的节奏和停顿；疏漏 4-9 个单词	有词汇、发音、句法或语言错误，但只是偶然发生
D (0-8)	能够朗读较短的句子；但有多处长时间的停顿	发音不准确，基本上不能理解	不能按照原语速朗读；疏漏 10 个以上单词	词汇、发音、句法或语言错误频发；影响理解

3.1.4 考生在朗读题型上的典型表述特点

考生在朗读题型上的典型表述特点有如下几点：

- 英语发音受母语发音影响严重，发音不准确。尤其是对于本文研究的广东省的中学生群体，受粤语母语的影响，大部分考生英语单词的发音不是很准确，对于生词，习惯于通过猜想发音。
- 朗读过程中跟不上字幕节奏，遗漏单词现象很普遍。在实际的考试过程中，字幕会按照视频的内容给出，每次停留的时间也会与原声持平，对于稍长的句子，大部分考生由于语速跟不上节奏，往往会出现还没朗读完，字幕已经消失或者出现下一句字幕，这种情况下，考生往往会直接漏掉最后几个单词，而进入下一句的朗读中。
- 遇到生词，造成停顿、重复以及自我修正等不流利现象。由于朗读题的文本中会出

现一定比例的生词、数字、缩写以及专业名词，考生在碰到这些词时，由于不确定发音，往往会出现停顿现象，或者自我修正发音，以及发音重复等现象。

- 语调平淡，基本不考虑韵律和重读。大部分考生忙于对字幕的跟读，很少有时间考虑句子的重读部分以及语调。

3.1.5 朗读题型自动评测的难点分析

基于上述对朗读题型的考察要点分析以及考生在朗读题型上典型表述特点的总结。朗读题型自动评测的难点分析如下：

- 克服非母语说话人母语的影响，提高语音识别系统的鲁棒性。一方面，内容的准确性是朗读题型评分的重要准则，另一方面，受母语的严重影响，语音识别系统的鲁棒性面临着比较大的挑战。因此，如何基于已有的语音识别自适应技术，努力提高语音识别系统的鲁棒性，是朗读题型自动评测的难点之一。
- 寻找具有高区分性的评分特征，区分测试者的口语水平。传统的文本相关评分特征通常选取流利度和发音质量的评分特征，对于本任务，如何根据任务的特点，提取既能表征口语水平，又具有高区分性的评分特征，也成为朗读题型自动评测的难点之一。

3.2 朗读题型数据库介绍

本文采用的朗读题型语音数据库来自 2011 年广东省肇庆市中考英语口语考试朗读题考生的真实数据，共收集 377 名考生数据，该语音库标记为 READ-DB。每位考生对应着一段语音，该语音为采样频率 16K 的 16Bit 单声道数据，长度为 1 分钟。我们邀请了具有英语教育背景的专家对数据进行了人工评分，评估的分数是整数，分为 0-20 分。READ-DB 的详细信息如下表 3.2 所示：

表 3.2 朗读题型 READ-DB 数据库信息

成绩等级	A	B	C	D
人数	149	31	40	157

3.3 朗读题型自动评分流程搭建

基于朗读题型的特点以及难点分析，参考已有朗读题型自动评分流程，我们提出了一套更符合当前任务的自动评分流程。下面将分别从语音识别系统、评分特征选择以及评分流程三个方面进行介绍。

3.3.1 语音识别系统搭建

本文采用的语音识别系统是基于剑桥大学开发的语音识别开源工具包 HTK 3.4.1 搭建而成。具体搭建包括特征提取、声学模型建模以及语言模型建模。

特征提取采用的是 MFCC_E_D_A 的 39 维声学特征，其中 MFCC 表示镁倒谱系数，

E 表示短时能量，D 表示一阶差分，A 表示二阶差分。

声学模型采用训练集为 TIMIT [32] 和 WSJ [33] 语料库训练而成的三音素模型。采用 CMU 音素集合，一共 40 个音素。每个音素包含 4 个高斯混合模型，其中静音模型包含 8 个高斯混合模型，声学模型共计 8000 个状态。

语音模型采用的是朗读题的文本，将文本进行语料预处理之后，训练基于词的二元文法语言模型。

词典由语言模型中所有词集构成。

3.3.2 评分特征

朗读题型自动评分通常使用的评分特征如下：

1）识别正确率和准确率

我们用 H 代表语音识别结果中正确的单词数，I 代表识别结果中插入的单词数，N_W代表参考文本的单词总数。用 Corr 和 Acc 分别代表识别正确率和准确率，则有：

$$Corr = \frac{H}{N_w} \times 100\% \tag{3-1}$$

$$Acc = \frac{H - I}{N_w} \times 100\% \tag{3-2}$$

这两维特征可以定量地考察考生发音的内容的完整性和准确性。

2）全局似然得分和局部似然得分

Viterbi 解码算法得到的识别结果的对数似然得分，很好的反映了母语发音和非母语发音的相似性，即说话人发音和声学模型之间相似程度。对于音素q_i，我们用d_i表示对应的持续时长，用l_i代表对应的对数似然得分，用 N 代表音素总个数，GL 代表全局似然得分，LL 代表局部似然得分，则有：

$$GL = \frac{\sum_{i=1}^{N} l_i}{\sum_{i=1}^{N} d_i} \tag{3-3}$$

$$LL = \frac{1}{N_p} \sum_{i=1}^{N} \frac{l_i}{d_i} \tag{3-4}$$

3）后验概率和 GOP [34]

给定发音文本$W = w_1, \ldots, w_M$，共有 M 个字（词），它可以被表示为音素序列$Q = q_1, \ldots, q_N$，共有 N 个音素。对于测试语音（即观测向量）$O = o_1, \ldots, o_T$，共有 T 帧。后验概率（Posterior Probability）是给定观测向量 O 的情况下音素序列 Q 的概率，也就是P(Q|O)。计算出的后验概率可以表示在观测到测试语音 O 后，对于候选音素序列 Q 的不确定性。该不确定性将被发音错误自动诊断系统用于判断 O 是否是音素 Q 的发音。

假设发音文本 W 所有的音素序列集合为ϕ，则根据贝叶斯公式，有音素的后验概率 PP 为：

$$\mathrm{PP} = \mathrm{P(Q|O)} = \frac{P(O|Q)P(Q)}{P(O)} = \frac{P(O|Q)P(Q)}{\sum_{q_i \epsilon \phi} P(O|q_i)P(q_i)} \tag{3-5}$$

通常情况下，音素序列集合 Q 中的元素对应的先验概率都被认为是相等的，则有：

$$\mathrm{P(Q|O)} \approx \frac{P(O|Q)}{\sum_{q_i \epsilon \phi} P(O|q_i)} \tag{3-6}$$

GOP（Goodness of Pronunciation）是后验概率的一种简化计算方法，该方法并不计算整段发音文本 W 的后验概率，而是利用单音素循环网络去计算发音文本所包含的每个音素各自的后验概率：对于属于$\mathrm{Q} = q_1, \ldots, q_N$的每一个音素$\mathrm{q}_i$，给定其对应的观测向量为 O，则可以用对数后验概率来表示该音素的发音质量，并做出如下假设，将(3-6)中分母的求和用最大值估计，这样就得到了 GOP 的定义：

$$GOP(p) = log\left(\frac{P(O|\mathrm{q}_i)}{\max\limits_{q_i \epsilon \phi} P(O|q_i)}\right) \tag{3-7}$$

GOP 分数可以通过两个过程得到：第一个过程使用将音频和正确的文本进行强制对齐，可以计算$P(O|q_i)$；第二个过程是单音素循环网络解码过程，与识别类似，首先需要构造一个音素循环网络作为识别词网络。例如，CMU 发音词典中共有 39 个音素，加上静音 sil 一共 40 个音素。由这 40 个音素构造一个自循环网络，然后利用 HTK 中的 HVite 命令进行音素识别，得到最有可能的音素序列，即$\max_{q_i \epsilon \phi} P(O|q_i)$。具体过程如图 3.2 所示：

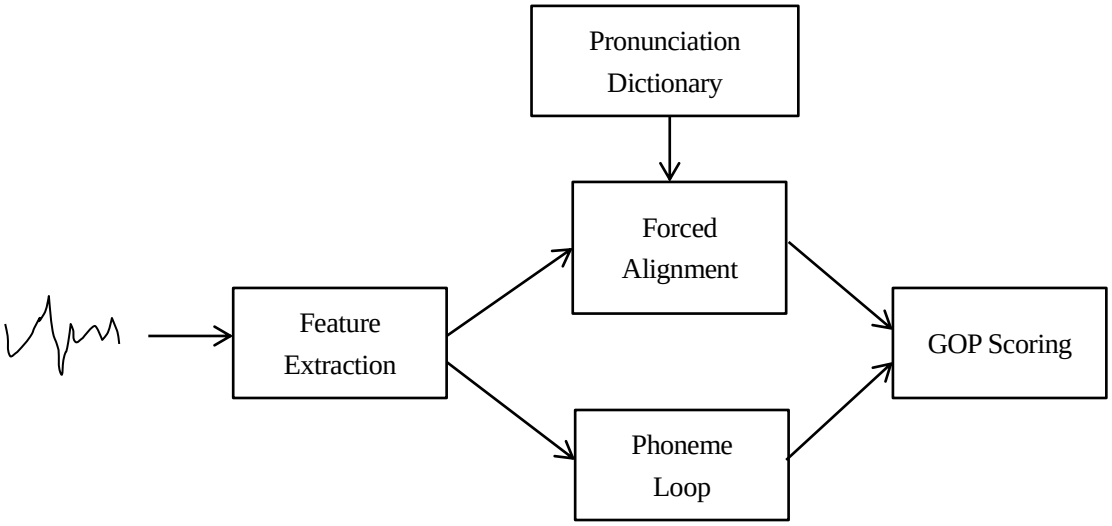

图 3. 2 GOP 的计算流程图

由图 3.2 可知：

$$GOP = S_{FA} - S_{PL} \tag{3-8}$$

其中S_{FA}就是强制对齐结果中每个音素每帧的维特比最大对数似然分数，我们用LP_{exit}代表音素结束处的对数似然分数，LP_{entry}代表音素开始处的对数似然分数，D_p代表音素的持续时长所对应的帧数，则有：

$$S_{FA} = \frac{LP_{exit} - LP_{entry}}{D_p} \tag{3-9}$$

S_{PL}，就是单音素循环网络解码结果中每个音素每帧的维特比最大对数似然分数。由于音素循环网络解码过程中产生的音素边界可能与强制切分不同，所以需要计算强制切分中覆盖的音素边界中的所有音素的加权平均帧的对数似然值。即：

$$S_{PL} = \frac{\sum_{i=1}^{m}(S_{(PL,i)}NF_{(PL,i)})}{D_p} \tag{3-10}$$

其中$NF_{(PL,i)} = f_e - f_s$，f_e和f_s代表每个覆盖音素的结束帧数和起始帧数，$S_{(PL,i)}$表示第 i 个音素模型的最大对数似然分数。

4）语速

在第二语言学习中，如中国人说英语的情况下，语速（Rate of Speech, ROS）能很好的表征说话的流畅度和口语熟练水平，对于初学者尤其如此。用 T 表示总发音时长，N 表示音素总个数，则有：

$$ROS = \frac{N}{T} \tag{3-11}$$

在具体评分中，考虑到非母语说话人辅音发音不够准确，可以只选取元音音素计算语速。

5）停顿时长比例和停顿数量

语音中词与词之间的短时停顿（Short Pause）也反映了发音流利度程度。一般说来，停顿时间越长，对应的流利度也就越差。我们用 SPR（Short Pause Ratio, SPR）代表短时停顿时长占总时长的比重，T 代表总的发音时长，t_i代表第 i 个停顿对应的时长，N_{sp}代表短时停顿的总个数，则有：

$$SPR = \frac{\sum_{i=1}^{N_{sp}} t_i}{T} \tag{3-12}$$

在自发性表述中，由于考生发音过程中存在大量的停顿，而每次停顿的时间可能很短，因此，停顿次数（Silence Number, SN）相比 SPR 可以更好地反映口语的流利程度。为了计算停顿数量 SN，需要在字典中为每个单词添加 sil 结尾的发音。

6）错误音素数量

音素发音检错技术也可以应用于发音自动评分中，如果能检测出音频中错误的音素数量（Error Phoneme Number, EPN），则也可以反映考生的发音质量。在本文中，我们采用的是基于 GOP 门限的初级发音检错技术，对每个音素q_i，取该音素的平均 GOP 值作为门限，低于 GOP 门限的音素视为错误音素发音，对每一个音素统计低于该音素 GOP 门限的音素数量，即可得到错误音素数量。假设错误音素数量为N_{error}，总音素数量为 N，则有错误音素数量 EPN 计算如下：

$$EPN = \frac{N_{error}}{N} \tag{3-13}$$

传统的文本相关的自动评分方法通常只选择反映音素质量和流利度的评分特征，例如 GL、LL、ROS 和 SPR。这是因为他们要处理的任务往往是句子级别的朗读，音频在内容上与标准文本没有什么差别。为了区分不同考生的口语水平，传统方法还会提取更精细的韵律和重音相关的评分特征。但是本文要处理的朗读题型，词数比较多，句子比较多，因而属于篇章级别的朗读。基于 3.1.4 节中对考生在朗读题型上的典型表述特点可知，考生在朗读过程中并不能保证内容都完整，因此内容相关的评分特征非常必要。而由于考生朗读的是篇章，韵律和重音方面的考察不是重点，反映韵律和重音的评分特征也没有足够的区分性。综合考虑这些因素以及评分特征的优劣，我们选取了如下 6 维评分特征，如表 3.3 所示：

表 3.3 朗读题型自动评分特征列表

特征名	描述	类别
GOP	Goodness of pronunciation，音素后验概率的近似	Intelligibility
EPN	Error phoneme number，音素错误数量	Intelligibility
ROS	Rate of speech in phoneme level，语速	Fluency
SPR	Short pause ratio，停顿时间占总语音时长的比例	Fluency
Corr	Correct rate of hypothesis，识别结果正确率	Content-related
Acc	Accuracy of hypothesis，识别结果准确率	Content-related

其中 GOP 和 EPN 属于表征音素发音质量的评分特征；ROS 和 SPR 属于表征考生发音流利程度的评分特征；Corr 和 Acc 属于表征单词级别内容准确性的评分特征。这些特征融合在一起，可以很好的区分考生的英语口语水平。

3.3.3 自动评分机制

3.3.3.1 评分流程

朗读题型评分流程分为预处理、特征提取和分数拟合三个阶段。其中预处理过程指的是利用语音识别系统对测试语音进行的一系列处理。传统的做法是依据朗读文本对测试语音进行强制对齐，匹配语音段的时间和置信度，作为特征提取部分的输入信息。

基于本文对朗读题型的特点和难点的分析可知，仅仅依靠强制对齐技术提取反映发音质量和流利度的评分特征，并不足以区分考生不同的口语水平。因此，本文提出了一

套新的评分流程，该流程可以提高语音识别的鲁棒性，获取更多反映考生口语水平的信息。其流程如图 3.3 所示：

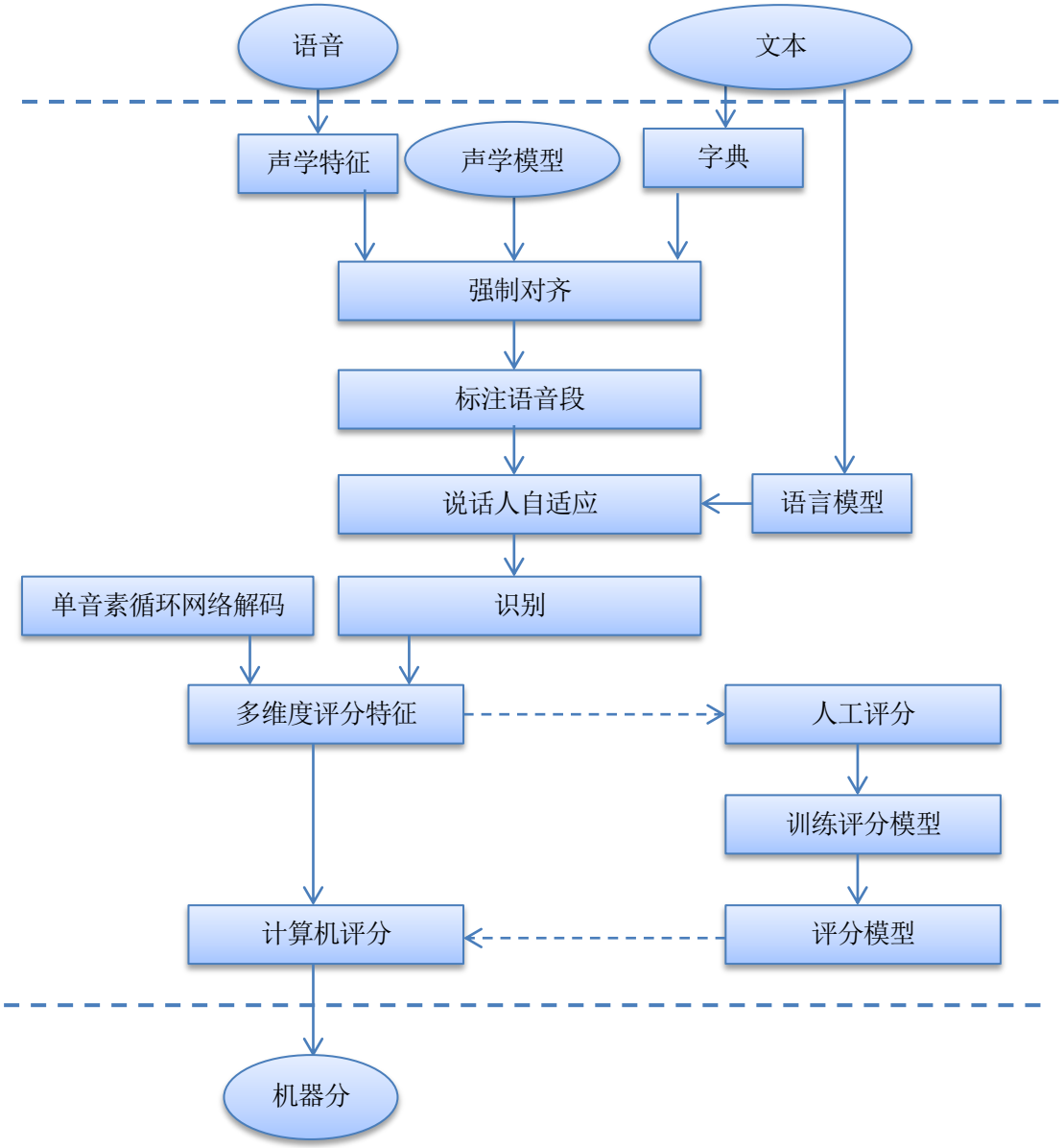

图 3. 3 朗读题型自动评分流程图

朗读题的自动评分基于语音识别技术，根据朗读文本在线训练语言模型，随后依次经过强制对齐（Forced Alignment）、标注语音段（Segmentation）、非监督性说话人自适应（Unsupervised Adaptation）、识别（Recognition）以及单音素循环网络解码（Phoneme Loop），然后从结果中提取需要的评分特征，包括识别 Corr、Acc、EPN、ROS、GOP、SPR 共计 6 维评分特征，将特征与老师评分进行多元线性回归（Linear Regression），得到评分模型，进行机器自动评分。

我们基于强制对齐获取考生语音段边界，并按照对齐结果中的 sil 标记对考生语音按句子级别进行标注时间段，这样做一方面可以提高后续两次识别的速度，同时也可以提高识别的精度。

为了提高非母语说话人语音识别的鲁棒性，我们采用了非监督性说话人自适应技术。首先，对语音进行一次语音识别，然后基于识别结果，利用 HERest 获取说话人语音相对于声学模型标准语音的转移矩阵，随后，利用转移矩阵信息，进行第二次语音识别，这种基于 MLLR 的非监督性说话人自适应技术可以有效地提高语音识别系统的鲁棒性。

单音素循环网络解码过程也是一个识别过程，不同的是这是音素级别的识别过程，并且没有任何语言模型的限制，采用音素循环网络进行自由解码，目的是获取最接近考生真实发音的音素序列以及对应的置信度信息，从而为计算 GOP 特征提供依据。

预处理阶段采用的语音识别系统如 3.3.1 节介绍所示。强制对齐和单音素循环网络解码过程使用的是单音素（monophone）声学模型，自适应及识别使用的是三音素（Tri-phone）声学模型。

特征提取阶段采用上节提到的 6 维评分特征，在此不再赘述。

3.3.3.2 分数预测

在目前流行的评分系统中，有多种进行评分模型训练及分数预测的方法，如线性回归（Linear Regression）、支持向量机（SVM）、人工神经网络（ANN）等方法。本文中，我们使用的是线性回归的方法。

线性回归是利用数理统计中的回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法之一，运用十分广泛。分析按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。

如果在回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

我们用 Y 表示因变量，X 表示自变量，简单的线性回归模型为：

$$\mathrm{Y}_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{3-14}$$

用$\mathbb{E}$表示均值，$\mathbb{V}$表示方差，则，对于上式，有：

$\mathbb{E}(\epsilon_i|X_i) = 0, \mathbb{V}(\epsilon_i|X_i) = \sigma^2$。模型中未知的参数为截距$\beta_0$，斜率$\beta_1$和方差$\sigma^2$。令$\hat{\beta}_0$，$\hat{\beta}_1$表示$\beta_0$，$\beta_1$的估计，拟合曲线为：

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3-15}$$

预测值或拟合值为$\hat{Y}_i = \hat{r}(X_i)$，残差定义为

$$\hat{\epsilon}_{\mathrm{i}} = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) \tag{3-16}$$

残差平方和或 RSS，衡量了曲线是否很好的拟合了数据，它定义为

$$\mathrm{RSS} = \sum_{i=1}^{n} {\hat{\epsilon}_i}^2 \tag{3-17}$$

一般来说，线性回归都可以通过最小二乘估计求出其方程。最小二乘估计就是使得 RSS 最小的$\hat{\beta}_0$和$\hat{\beta}_1$值。最小二乘估计为：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} \tag{3-18}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n \tag{3-19}$$

σ^2的无偏估计为

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n} \hat{\epsilon}_i{}^2 \tag{3-20}$$

本文中，令评分模型$\vec{\mathrm{A}}$是一维矢量，可理解为各位评分特征$\vec{F}$的权重。在训练集上，我们已知人工评分S_H，根据下面的等式，计算$\vec{\mathrm{A}}$和其偏移量 b，使其在最小均方误差意义下最优。

$$S_H = \vec{F} \cdot \vec{A} + b \tag{3-21}$$

在测试集上，我们就可以利用训练集训练的最优评分模型$\vec{A}_{opt}$和b_{opt}计算机器分S_M，如下式：

$$\mathrm{S}_M = \vec{F} \cdot \vec{A}_{opt} + b_{opt} \tag{3-22}$$

在实验中，为了验证机器分的准确性，通常在训练集中采用 K 重交叉验证，它的算法如下：

（1） 将数据随机分成 K 个大小近似相等的部分。

（2） 对于$k = 1$到 K，执行下列步骤：

（a） 将第 k 个部分从数据中删除。

（b） 由余下的数据计算分类器$\hat{\mathrm{h}}_{(k)}$。

用$\hat{\mathrm{h}}_{(k)}$来预测第 k 个部分中的数据[35]。

3.4 朗读题型实验与结果分析

朗读题型自动评测的实验在 READ-DB 数据库上进行，采用 READ-DB 及对应的人工评分训练评分模型，采用留一交叉验证预测测试音频的机器得分。本文评分模型的训练以及机器分的预测都在 R 软件下完成[36]。

发音评分系统的评分本质就是计算机模拟人工进行打分的过程，系统性能的好坏，是和真实人工评分对比的结果。在本文中，我们用相关性和平均分差来衡量自动评分的性能。

- 相关性：测试集上指机器评分和人工评分的相关性。相关性衡量了机器分和人工评分在线性意义上的相似度，是评价系统的重要指标。我们用S_M表示机器分，S_H表示人工评分，用σ_M表示机器分的方差，σ_H表示人工评分的方差，$\rho_{H,M}$表示机器分与人工评分的相关度，Cov表示协方差运算，其计算公式为：

$$\rho_{H,M} = \frac{Cov(S_M, S_H)}{\sigma_M \sigma_H} \tag{3-23}$$

- 平均分差：测试集上机器分和人工评分之差的绝对值的平均，反映的机器分在具体评分上和人工评分的偏离程度。平均分差一般也简称为分差，用d表示，用$\mathbb{E}$表示均值运算，则有：

$$d = \mathbb{E}|S_M - S_H| \tag{3-24}$$

3.4.1 实验结果

为了增加实验结果的可比性，我们同时采用了传统基于强制对齐的朗读题自动评分方法，选用的评分特征集为 GL、LL、ROS 和 SPR，由于传统方法仅仅进行强制对齐，因此所有特征的计算都基于强制对齐的结果。其中实验结果如表 3.4~3.6 所示。

表 3. 4 READ-DB 上传统评分流程的评分特征与人工评分相关系数

评分特征	GL	LL	ROS	SPR
相关系数	0.702	0.373	0.525	-0.665

表 3. 5 READ-DB 上新评分流程的评分特征与人工评分相关系数

评分特征	GOP	EPN	ROS	SPR	CORR	Accuracy
相关系数	0.763	-0.796	0.908	-0.719	0.870	0.730

表 3. 6 READ-DB 上传统方法与新方法的自动评分性能对比

	相关系数	平均绝对分差
传统评分流程	0.782	3.81
新的评分流程	0.923	2.32

3.4.2 结果分析

对比表 3.4 和表 3.5，本文采用的评分特征具有更好的区分性。表 3.6 说明，与传统特征相比，本文提出的朗读题型自动评分流程具有更高的机器分与人工评分相关度，这充分说明本文提出的基于强制对齐、标注语音段、非监督性说话人自适应、语音识别以及单音素循环网络解码评分流程的必要性，以及提取的基于发音质量、流利度和内容相关评分特征的有效性。

3.5 朗读题型自动评测的改进措施

除了在 READ-DB 数据库上进行朗读题型自动评测的实验，笔者在就读期间还处理累计 80000 人次的广东省英语中考朗读题自动评测任务，在实际的评分过程中积累了如下有效的改进措施：

- 基于音频预处理的改进措施

在大规模海量数据处理时，音频预处理显得非常有必要。由于本文应对的朗读题型自动评测任务的音频为考生真实的音频数据，而非实验室环境下的录制数据，因此音频具有如下特点：首先，音频录音时长为 1 分钟，由于考生跟随字幕进行朗读，因此在音频的开始段和结束段存在着大量的静音段，严重时，静音段比例达到了 50%。另一方面，由于考生在教室里进行作答，由于设备不统一，环境噪声等影响，大量考生音频存在着零点漂移现象，并且音频开始和结束的静音段中也存在着由于环境噪声、喘息声等非语音噪声。为了应对这些特点，本文在实际的评测中，增加音频预处理过程，主要包括去除直流分量操作和切除首尾长的静音段操作，经过音频预处理操作的语音，解码时间可以缩短三分之一，解码精度也可以提高。

- 基于扩充评分模型的改进措施

由于本文采用的朗读题型自动评测方法是一个典型的机器学习的过程，评分模型的好坏的对评分结果有着重要的影响。尤其是训练数据中分数分布的均衡性对模型影响很大，如果训练数据中高分段过多，低分段数据不足，训练出来的评分模型可能会导致机器评分对低分段数据评分过高的现象，反之亦然。因此，当评分模型表现欠佳时，可以考虑增加部分训练数据，平衡训练数据的不同分数段的数量，达到扩充评分模型，提高评分精度的目的。

3.6 小结

本章从朗读题型的特点和难点分析开始，全面讨论了朗读题型的考察要点、评分标准、考生在朗读题型上的典型表述特点以及由此带来的自动评测的难点。随后基于这些特点和难点，提出了基于强制对齐、标记语音段、非监督性说话人自适应、语音识别以及单音素循环网络解码操作的自动评分流程，提取了反映发音质量、流利度和内容相关的评分特征集（GOP、EPN、ROS、SPR、Corr 和 Acc），并在 READ-DB 数据库上进行了实验，实验结果表明，相比于传统基于强制对齐的评分方法，本文提出的评分流程具有更高的机器评分与人工评分的相关度。最后，基于大规模海量数据的朗读题自动评测经验，本文提出了基于音频预处理及扩充评分模型的改进措施。

第四章 复述题发音自动评测技术

本章将研究复述题型的自动评测技术。我们首先基于复述题型的特点，重点分析考生在复述题型上的典型表述特点，随后搭建复述题型的自动评分流程，并提出内容相关的评分特征，在数据库上进行实验。随后，我们将重点放在复述风格语言模型的构建上，提出了针对复述题的基于混合语言模型的语言模型自适应方法。

4.1 复述题型特点及难点分析

4.1.1 复述题型简介

本文所研究的复述题型是基于听力的复述题型，其考察方式如下：系统播放一段英文音频（无字幕），音频内容通常是描述一个小故事，200~300 词，然后要求考生根据听到的内容进行复述，复述可以与音频内容完全一样，也可以按自己的方式表达，但是基本内容要求与音频一样。回答时长 2 分钟。题目会给简短的中文提示以及单词提示，图 4.1 是复述题型示例。

与朗读题相比，复述题的参考文本只起参考作用，该类题型允许考生真实发音与范文有很大的差距。可见它不仅考察了考生的发音质量，还考察了其语言组织能力，是一种更能体现英语口语水平和思维能力的题型。

4.1.2 复述题型考察要点

本文研究的复述题型属于基于听力的复述题型，首先要考察的是考生的听力能力以及对听到的内容的理解能力，反映在考生的回答上，则是考生的回答与范文之间内容的相关性，相关性比较高，说明考生听懂了；其次，考察的是考生的组织语言的能力，具体包括考生在复述过程中使用的句式、语法以及词汇等方面是否正确；最后，还要考察考生的语速、停顿、发音以及语调等口语能力。由此可以看出，复述题型相比于朗读题型，难度更大，考察得更全面，更能反映考生真实的口语水平。

4.1.3 复述题型评分标准

基于上节提到的考察要点，朗读题评分标准需要综合考虑考生答案的流利度（Fluency）、连贯（Coherence）、发音（Pronunciation）、内容（Content）、准确度（Accuracy），基于各项内容加权表征考生的口语水平。总分为 24 分，分为 ABCD 四个等级。具体评分细则如表 4.1 所示：

梗概：
一位壮汉在木工厂找到一份砍树的工作。第一天他干得非常好。他想好好干。可是他的工作成效却一天比一天差。后来他知道了原因。

关键词：
axe（斧头）、cut down (砍倒)、sharpen（使变得锋利）、dull（迟钝的）

故事录音文字：
Once upon a time a very strong man asked for a job in a wood factory, and he got it.[1] His boss gave him an axe and showed him the area where he was to work and the first day the woodcutter cut down 18 trees. [2]
'Congratulations,' the boss said. 'Go on that way!'
Very excited for the boss' words, the man tried harder the next day, but he could only cut down 15 trees. The third day he tried even harder, but he could only cut down 10 trees. Day after day he was cutting fewer and fewer trees. [3]
'I must be losing my strength,' the man thought. He went to the boss and apologized, saying that he could not understand what was going on.
'When was the last time you sharpened your axe? ', the boss asked.
'Sharpen! I had no time to sharpen my axe. I have been very busy trying to cut down trees.'[4]
Our lives are like that. We sometimes get so busy that we don't take time to sharpen the 'axe'.
In today's world, it seems that everyone is busier than ever, but less happy than ever. Why is that? Could it be that we have forgotten how to stay 'sharp'?[5] （207 words）

图 4.1 复述题型示例

表 4.1 复述题型评分标准

Feature / Grade	流利度，连贯 (Fluency & Coherence)	发音 (Pronunciation)	内容 (Content)	准确度 (Accuracy)
A (20-24)	表达流畅，连贯，自然；逻辑清晰	重音和音韵准确；利用不同语调表达意义；单独词汇发音准确，容易被理解	能够清晰地、完整地复述原文中的 5 个要点	几乎没有词汇、发音、句法或语言错误
B (15-19)	表达流畅，连贯，自然；逻辑清晰，但有个别处影响自然语流	发音准确，利用重音和语调，但有一些语法错误	基本清晰地、完整地复述原文至多疏漏 1 个要点	只有极少量的词汇、发音、句法或语言错误
C (9-14)	尽管有些较长的停顿，但能够有节奏地连贯表达较长的句子	母语音较重，导致一些重音和语调使用不当；但不影响理解	表达了原文的主题但疏漏 2-3 个要点	有词汇、发音、句法或语言错误，但只是偶然发生
D (0-8)	能够使用较短的句子;但应表达困难，有多处长时间的停顿	发音不准确，基本上不能理解	不能完整表达原文；疏漏 4 个以上要点	词汇、发音、句法或语言错误频发；影响理解

4.1.4 考生在复述题型中的典型表述特点

考生在复述题型中的表述属于非母语说话人自发性表述，有着许多典型的表述特点，本文在第二章中也曾对非母语说话人在英语口语表述上的特点做过调研，在此，基于考生的实际音频，总结考生在复述题型中的典型表述特点，分为语音学特点和语言学特点。

其中语音学方面的特点包括：

- 英语发音受母语发音影响严重，发音不准确。这一点和朗读题型的表述特点类似，尤其是对于本文研究的广东省中学生的英语口语，受粤语影响更明显。
- 固定的错误的发音习惯。例如，喜欢在单词后面加/s/或者/z/等尾音。
- 下意识造词。由于考生复述过程只记得单词词根，于是在表述该词的衍生词时下意识造词。比如在形容词后面加 ness 构成名词，在动词后面加 ing 构成现在分词等。

例如下面的句子：

(1) Although life is full of *beater and beatenness* and sweet, john knew a truth sorrow and joy came together at the same time and love last forever.

(2) A lot of surgery had *beening* done on her, but all was in vain.

在这两句话中，beater、beatenness、beening 都是不存在的词，但是考生下意识制造了出来。

语言学方面的特点包括：

- 话题与原文故事一致，但是用词和句式上存在变化。由于考生的复述是基于听到的原文故事进行展开，考生的表述内容在话题上会与原文紧密相关；但由于考生也可以使用自己的语言进行表述，因此在用词上又存在诸多变化。
- 表述中存在大量不流利现象（停顿、重复和自我修复等）。由于考生需要自己组织语言进行表述，因此具备第二章提到的非母语说话人英语口语表述的一些典型特点。
- 表述中存在不少词法和句法方面的表述错误。由于考生的英语口语水平相比母语说话人有很大的距离，因此存在乱用时态、单复数不分以及一些错误的句法表述习惯。

4.1.5 复述题型自动评测的难点分析

基于上述对考生在复述题型中的典型表述特点，可以看出相比于朗读题型，复述题型自动评测难度更大，主要面临两个问题：

- 选择合适的语料构建复述风格的语言模型，提高非母语说话人自发性表述的识别率。在复述题中，考生在表述内容与范文紧密相关，但在用词上又具有诸多变化，因此不能简单地按照朗读题型那样仅使用原文故事文本进行构建语言模型。同时，作为非母语说话人自发性表述，考生的表述中存在大量的不流利现象，并且包含不少的词法和句法方面的表述错误。这说明传统的电话对话转写文本以及新闻文本也不适合用来构建复述题的语言模型。这些典型的复述风格，为语音识别过程中的语言模型建模以及识别率带来了很大的困难。也就是说，如何提高非母语说话人自发性表

述的识别率问题，成为复述题自动评测需要解决的一个难点。

- 在识别率不高的情况下，寻找高区分性的内容相关的评分特征。一方面，与朗读题型自动评测不同，复述题型的评分标准虽然也要考虑发音、语速、语调、停顿等因素，但是更倾向于考生回答与参考答案在内容上的相似度。另一方面，复述题型中语音识别率并不高。因此，如何克服不高的识别率，寻找基于内容相关的评分特征，成为复述题型自动评测过程中需要解决的另一个难点。

4.2 复述题型数据库介绍

本文采用的朗读题型语音数据库为 2010 年广东省深圳市中考英语口语考试复述题考生真实音频数据，共收集 279 名考生数据，该语音库标记为 RETELL-DB。每位考生对应着一段语音，该语音为采样频率 16K 的 16Bit 单声道数据，长度为 2 分钟。我们邀请了具有英语教育背景的专家对数据进行了人工评分，评估的分数是整数，分为 0-24 分。RETELL-DB 的详细信息如表 4.2 所示：

表 4.2 复述题型数据库信息

成绩等级	A	B	C	D
人数	27	49	88	115

4.3 复述题型自动评分流程搭建

基于复述题型的特点以及难点分析，我们将重点放在提取内容相关的评分特征以及构建复述风格的语言模型两个难点问题上。在这一节，我们重点放在第一个问题上，并将分别从语音识别系统、评分特征选择以及评分流程三个方面介绍复述题型自动评分流程。

4.3.1 语音识别系统搭建

与朗读题型搭建的语音识别系统类似，复述题型自动评测中搭建的语音识别系统也建立在 HTK 基础之上。特征提取采用的是 MFCC_E_D_A 的 39 维声学特征。 声学模型采用训练集为 TIMIT 和 WSJ 语料库训练而成的三音素模型。采用 CMU 音素集合，一共 40 个音素。每个音素包含 4 个高斯混合模型，其中静音模型包含 8 个高斯混合模型，声学模型共计 8000 个状态。语言模型采用的是范文语料以及相关的口语语料构建的语言模型。

4.3.2 评分特征

为了更好地考察考生回答与参考答案之间在内容上的相关度，我们提出了 Similarity（内容相似度）、KCR（关键词覆盖率）、WN（单词数）、UWN（非重复单词数）四维内容相关的评分特征，并结合传统的 GL（全局似然度）、LL（局部似然度）反映发

音质量的评分特征以及 ROS（语速）、SN（停顿次数）反映流利度的评分特征，共同构成复述题型自动评测的评分特征。

在这里，首先介绍四维内容相关的评分特征。

- Similarity（内容相似度）：由于复述题型中考生可以按照自己的方式进行复述，复述题型的参考答案并不唯一。因此，识别结果与多种可能的参考答案之间的相似度的最大值更能准确地体现考生回答与参考答案之间的内容相似度。我们用S_i表示识别结果与第 i 个参考答案之间的内容相似度，共有 n 个参考答案，则有

$$Similarity = \max(S_1, S_2, \dots S_i, \dots S_n) \tag{4-1}$$

我们通过基于动态规划的字符串对齐来计算识别结果与第 i 个参考答案之间的内容相似度。

- KCR（关键词覆盖率）：我们首先基于对范文的分析，得出考生复述过程中应该提及的关键词集合（关键词未必一定包含在原始范文中），假设共有 n 个关键词，用k_i表示关键词集合中第 i 个关键词。则有

$$KCR = \frac{\sum_{i=1}^{n} cover(k_i)}{n} \tag{4-2}$$

其中，

$$\text{cover}(k_i) = \begin{cases} 1, if\ k_i \text{ occured in the speech} \\ 0, if\ k_i\ didn'toccur\ in\ the\ speech \end{cases} \tag{4-3}$$

- WN（单词数）、UWN（非重复单词数）：一方面，通常情况下，如果考生在复述过程中使用的单词比较多，考生表达的内容也比较丰富。因此，识别结果中单词的总数量（Word Number, WN）也可以反映考生回答的内容充实性。另一方面，如果考生回答中重复的单词太多，则包含的有用信息太少，因此识别结果中非重复的单词数目（Unique Word Number, UWN）可以反映考生回答中包含的有用信息量以及用词的广泛度。

我们可以用图 4.2 表示提取内容相关的评分特征的流程图。我们首先基于原文故事扩展可能的参考答案表述，得到参考答案 1~n。在本文的实验中，我们采取的是基于规则的自动扩展与手动扩展相结合的半自动方法。具体步骤如下：

1）通过词性分析，提取原文故事中的动词、名词，组成关键词集。

2）对动词进行不同时态的变换，确定动词的扩展集；对名词进行单复数变换，确定名词的扩展集；得到扩展的关键词集。

3）用扩展集中的单词替换原文故事中的对应单词；

4）人工对自动扩展的结果进行句式、语义上的调整，得到扩展的参考答案。

随后基于 ASR 的结果，我们将识别结果与扩展出来的参考答案一一进行内容相似度计算以及关键词覆盖率的计算，得到上面提出的 4 维内容相关的评分特征集。

结合传统的反映发音质量和流利度的评分特征，我们得到复述题型自动评测过程中使用的评分特征集，如表 4.3 所示：

表 4.3 复述题型自动评分特征列表

评分特征	描述	类别
Similarity	识别结果与多种可能的参考答案之间的相似度的最大值	Content-related
KCR	Keyword coverage rate，关键词覆盖率	Content-related
WN	Number of words，识别结果中单词数目	Content-related
UWN	Number of unique words，识别结果中不重复单词数目	Content-related
GL	Global log-likelihood，识别结果的全局似然得分	Intelligibility
LL	Local log-likelihood，识别结果的局部似然得分	Intelligibility
ROS	Rate of speech in phoneme level，音素级别语速	Fluency
SN	Number of silences，识别结果中停顿数量	Fluency

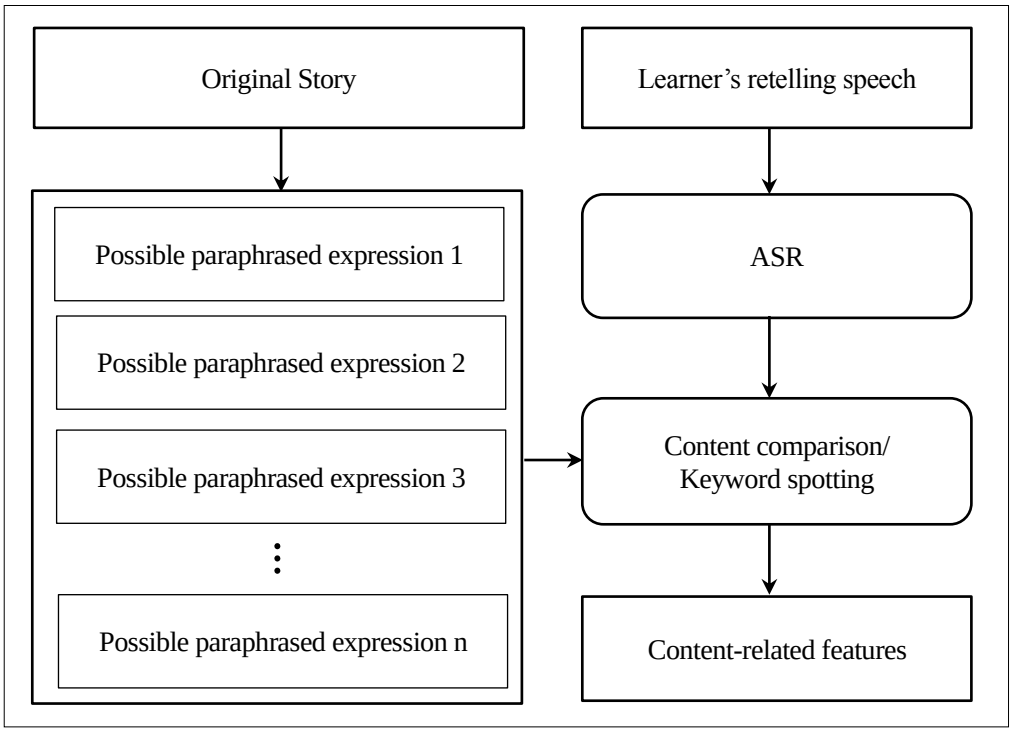

图 4.2 复述题型内容相关评分特征提取流程图

4.3.3 自动评分机制

复述题自动评分流程与朗读题类似，也分为预处理、特征提取和分数拟合三个阶段具体流程图如图 4.3 所示。

复述题自动评分系统采用范文和口语语料在线训练语言模型，先后进行非监督性说话人自适应（Unsupervised Adaptation）及识别（Recognition），从识别结果中提取上节提到的 8 维评分特征。然后将评分特征与老师评分进行多元线性回归（Linear Regression），得到评分模型，进行机器自动评分。

其中，非监督性说话人自适应、识别均在 HTK 下完成，声学模型与朗读题型所使用的声学模型相同。

4.4 复述题型实验结果与分析

复述题型自动评测的实验在 RETELL-DB 数据库上进行，与朗读题型自动评测实验类似，采用留一交叉验证，用机器评分与人工评分结果的相关系数来衡量自动评分的性能。

4.4.1 实验结果

为了增加实验结果的可比性，我们将对比了传统评分特征集合（GL、LL、ROS、SN）与本文提出的基于内容相关评分特征集合（Similarity、KCR、WN、UWN）的实验结果，实验结果如表 4.4~4.5 所示：

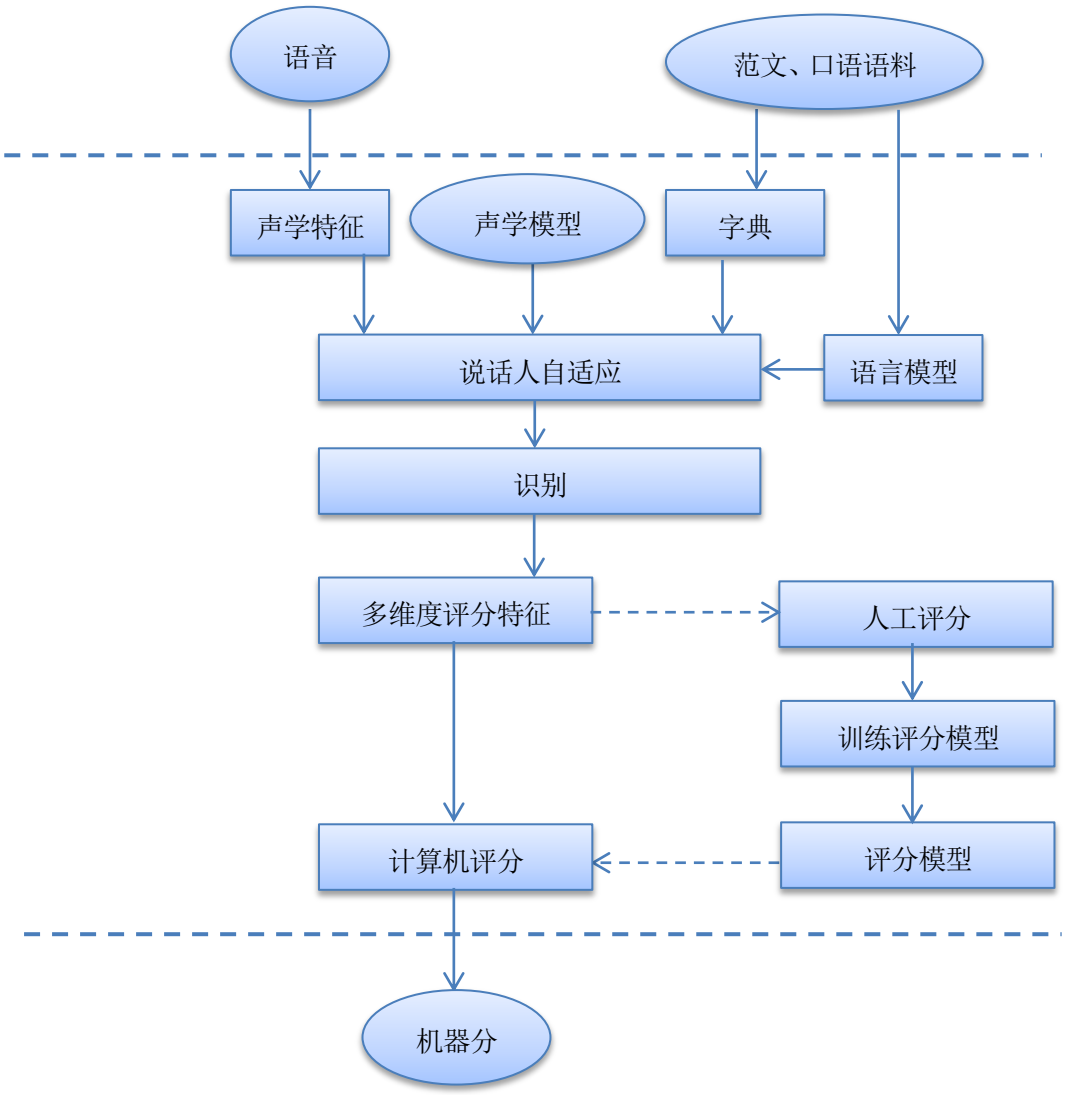

图 4. 3 复述题型自动评分流程图

表 4.4 RETELL-DB 数据库评分特征与人工评分的相关系数

评分特征	Similarity	KCR	WN	UWN	GL	LL	ROS	SN
相关度	0.462	0.496	0.285	0.481	0.231	0.306	0.328	-0.148

表 4.5 不同评分特征集在 RETELL-DB 自动评分性能对比

	相关系数	平均绝对分差
GL,LL,ROS,SN	0.484	4.25
GL,LL,ROS,SN, Similarity, KCR,WN,UWN	0.621	3.68

4.4.2 结果分析

从表 4.4 可以看出，相比于传统基于发音质量和流利度的评分特征，本文提出的基于内容相关的评分特征与人工评分具有更好的相关性。表 4.5 说明，增加了本文提出的内容相关的评分特征集合之后，自动评分与人工评分的相关度由 0.484 提高到 0.621，这充分说明内容相关评分特征的有效性。

但是，与朗读题自动评测相比，复述题型自动评测的机器分与人工评分的相关性还有很大的改进空间，这是因为复述题型中面临的非母语说话人自发性表述的语音识别率还有待提高。下一节，将具体介绍复述风格语言模型的构建方法。

4.5 复述风格语言模型构建方法

本节将探索复述题型自动评测的第二个难点问题——提高非母语说话人自发性表述的语音识别率，重点从语言模型的角度进行研究。接下来，将详细介绍文本语料的收集、基于混合语言模型的语言模型自适应方法以及实验。

4.5.1 收集文本语料

为了构建复述风格的语言模型，首先需要搜集相关的文本语料。但是，由于人工转写复述题音频需要花费巨大的人力和较长的时间，因此，完全通过转写获得匹配的复述风格的文本语料并不现实。我们只能基于考生在复述题型中的典型表述特点，收集不同的文本语料进行组合，达到复述风格文本语料的效果。

在 4.1.5 节，我们曾从语言学的角度分析了考生在复述题型中的典型表述特点。现在，我们将基于这三个典型的特点，选择和收集文本语料。首先，我们将选择复述题原文故事文本作为话题相关的文本语料（Topic-related Corpus，TR），因为考生的表述是基于听到的原文故事展开的，所以，原文故事文本可以提供大量的话题相关的词汇。但是，由于考生在复述过程也可以使用自己的语言，仅仅使用原文故事文本并不能覆盖考生复述过程可能出现的词汇。因此，我们还将搜集中学生英语学习过程中的一些文本资料，我们称之为书面风格的文本语料（Document Style related Corpus，DS）。这部分文本语料包括中学生英语教材、英语口语教程、新概念英语教程以及历年的英语中高考试题，可见，这部分文本语料可以覆盖中学生的词汇范围。此外，考生的表述属于非母语说话人自发性表述，一方面包含了大量的口语表述的典型的不流利现象（停顿、重复和自我修复等）；

另一方面又包含了可能的词法和句法错误，因此，我们还需要口语风格的文本语料（Spoken-style related Corpus，SS）。在此，我们选择了《中国学生英语口笔语料库 SECCL 2.0》[37]，并采用了其中复述题的转写文本部分，共包含 2003-2006 年共计 713 名英语专业二年级学生在全国英语专业四级考试中的复述题转写文本（4 个话题）。表 4.6 是收集的文本语料的详细情况列表：

表 4.6 文本语料的详细情况列表

Corpus	TR	SS	DS
#Words	537	159.1K	275.6K
#Uniq. words	212	2616	11712
#Sentences	42	8268	20095

4.5.2 基于混合语言模型的语言模型自适应方法

传统的混合语言模型（Mixture of Language Model）是通过多个语言模型线性插值（linear interpolation）实现的。插值的对象可以是基于词的语言模型（Word-based N-gram Language Model），也可以是基于词的语言模型与基于词类的语言模型（Class-based Language Model）。我们以 3 元文法为例，设共有 J 个独立的基于词的语言模型，则插值后的语言模型，词的概率计算如公式（4-4）所示：

$$\mathrm{P}(w_i|w_{i-1},w_{i-2}) = \sum_{j\epsilon J} \lambda_j P_j(w_i|w_{i-1},w_{i-2}) \tag{4-4}$$

其中，插值系数λ_j可以通过期望最大值算法（Expectation-Maximization algorithm）在一部分预留的测试集上进行优化求出。并且有如下约束：

$$\sum\nolimits_{j\in J} \lambda_j = 1 \tag{4-5}$$

对于基于词的语言模型和基于词类的语言模型的插值，有类似的公式：

$$\mathrm{P}(w_i|w_{i-1},w_{i-2}) = \sum_{j\epsilon J} \lambda_j P_j(C_i|C_{i-1},C_{i-2}) P(w_i|C_i) \tag{4-6}$$

其中C_i代表词类。

传统基于混合语言模型的语言模型自适应方法是简单将所有文本语料混合在一起，分别训练基于词的语言模型和基于词类的语言模型，然后将两种语言模型进行插值，得到自适应语言模型。在此，文本提出了一种新的基于混合语言模型的语言模型自适应方法，并将其应用于复述风格的语言模型构建中。

我们首先分别对 TR 文本语料、SS 文本语料和 DS 文本语料训练基于词的语言模型和基于词类的语言模型。然后对基于词的语言模型进行插值，得到基于词的混合语言模型，对基于词类的语言模型进行插值，得到基于词类的混合语言模型。最后，我们将基

于词的混合语言模型和基于词类的混合语言模型进行插值，得到自适应语言模型，也就是复述风格的语言模型。整个方法的流程图如图 4.4 所示。

4.5.3 实验结果

我们采用 4.5.1 节中提到的话题相关文本语料（TR Corpus）、口语风格文本语料（SS Corpus）和书面风格文本语料（DS Corpus）作为语言模型的训练集。为了对自适应语言模型进行评估，我们将采用真实的复述题型转写文本作为测试集。我们邀请具有英语教育背景的专家分别对西乡中学和育才中学考生的复述题的真实音频进行转写，得到了 XIXIANG 和 YUCAI 两个测试集，两个测试集包含了两个不同的话题。其中，XIXIANG 测试集是 280 名考生真实复述音频的转写文本，包含了 2750 个句子，共计 28150 个词；YUCAI 测试集是 15 名考生真实复述音频的转写文本，包含了 207 个句子，共计 2257 个词。

按照 4.5.2 节提出的语言模型自适应流程，我们采用 SRILM [38]训练了复述风格的语言模型。其中，图 4.4 中 Topic-related language model 和所有的 class-based language model 都采用了 Witten-Bell 平滑算法[39]进行平滑；Spoken-style N-gram language model 和 Document-style language model 采用了 interpolated modified Kneser-Ney 平滑算法[40]进行平滑。对于每种语言模型，我们分别训练了二元、三元和四元的语言模型，其中 cutoffs 均为 1。所有的插值系数，我们都使用了 SRILM 中的 compute-best-mix 工具进行优化。

考虑到高中生的词汇量（低于 5000）和训练集的大小，我们针对不同的训练集选取了不同的词汇集。因为考生的回答是从听到的复述题故事原文中衍生出来的，所以我们选用了 TR 文本语料中所有的词汇来训练话题相关的语言模型（topic-related language models）。对于 SS 文本语料和 DS 语料，我们基于经验分别选取了词频最高的 2500 词和 3500 词，用来训练口语风格的语言模型（spoken-style language models）和书面风格的语言模型（document-style language models）。因此，通过线性插值，基于词的混合语言模型（N-gram Mix language model, N-M-LM）、基于词类的混合语言模型（Class-based Mix language model, C-M-LM）以及自适应语言模型（Adapted language model, A-LM）的词汇表均共有 4373 个词。此外，我们分别将词类选为 50，200 和 300，采用 full greedy merging algorithm，针对 TR、SS 和 DS 三个文本语料库分别训练词类语言模型。

为了与传统的基于混合模型的语言模型建模方法作比较，我们先将 TR、SS、和 DS 三个文本语料库进行合并，随后在合并的语料上训练基于词的语言模型（N-gram language model, N-LM）和基于词类的语言模型（Class-based language model, C-LM），并对两者进行线性插值，得到混合语言模型（Mixture language model, M-LM）。其中，对于词类语言模型，我们基于经验将聚类数量选为 500。

我们在两个测试集上对所有的混合语言模型进行评估，评估标准为困惑度。由于所有的混合语言模型的词汇集都是一样的，因此他们的 OOV（Out Of Vocabulary）在相同的测试集上也都一样，其中 XIXIANG 测试集上的 OOV 为 1.18%，YUCAI 测试集上的

OOV 为 1.24%。表 4.7~4.8 分别是所有的混合语言模型分别在 XIXIANG 和 YUCAI 测试集上的困惑度列表；图 4.5~4.6 更直观的显示了不同的混合语言模型在两个测试集上困惑度的差别。

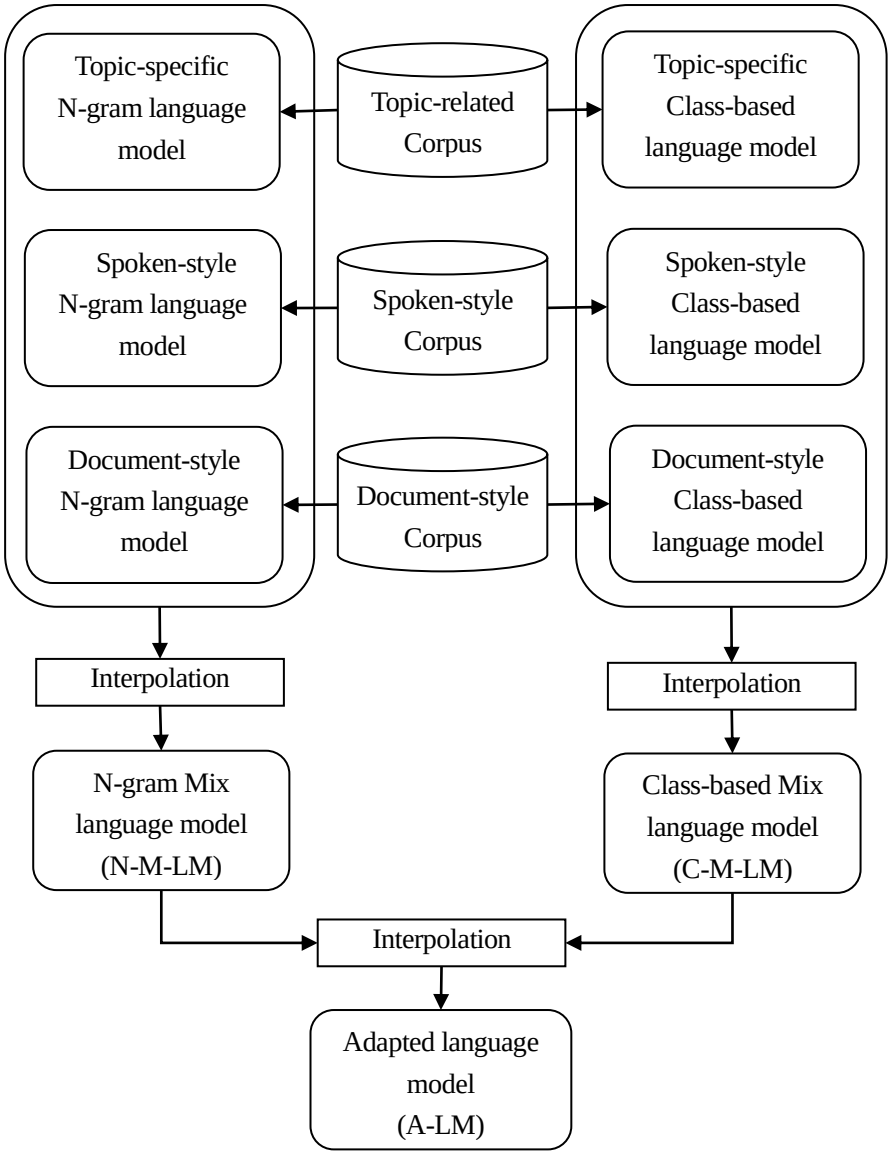

图 4. 4 基于混合语言模型的语言模型自适应方法流程图

表 4. 7 XIXIANG 测试集上混合语言模型的困惑度

LM	2-gram	3-gram	4-gram
N-LM	153.586	138.968	134.119
N-M-LM	82.735	71.961	71.543
C-LM	158.655	162.104	167.211
C-M-LM	79.783	77.199	77.199
M-LM	146.506	132.831	129.529
A-LM	77.334	69.684	69.336

表 4. 8 YUCAI 测试集上混合语言模型的困惑度

LM	2-gram	3-gram	4-gram
N-LM	154.888	111.921	106.872
N-M-LM	60.776	43.508	43.025
C-LM	176.971	150.592	158.884
C-M-LM	52.891	49.477	49.036
M-LM	149.186	107.461	102.981
A-LM	51.593	42.785	42.152

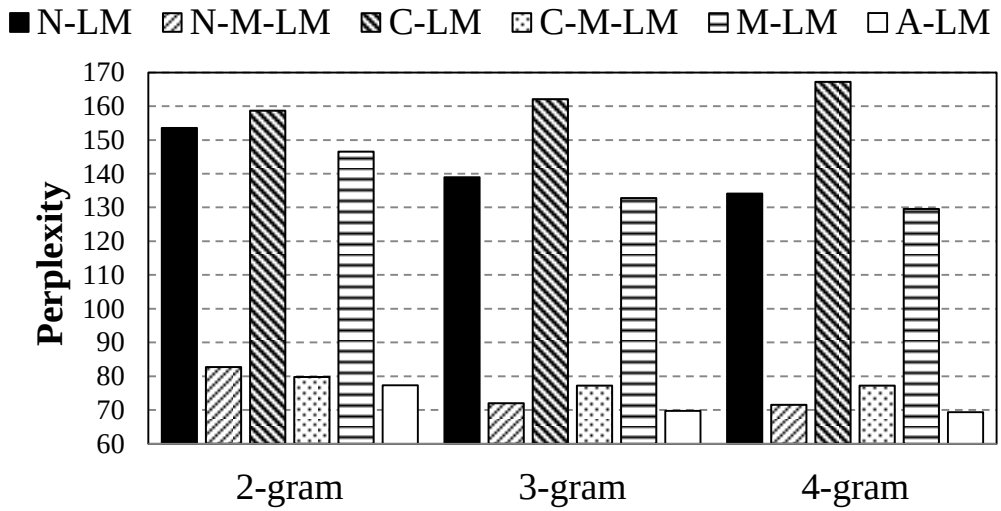

图 4. 5 XIXIANG 测试集上混合语言模型的困惑度对比

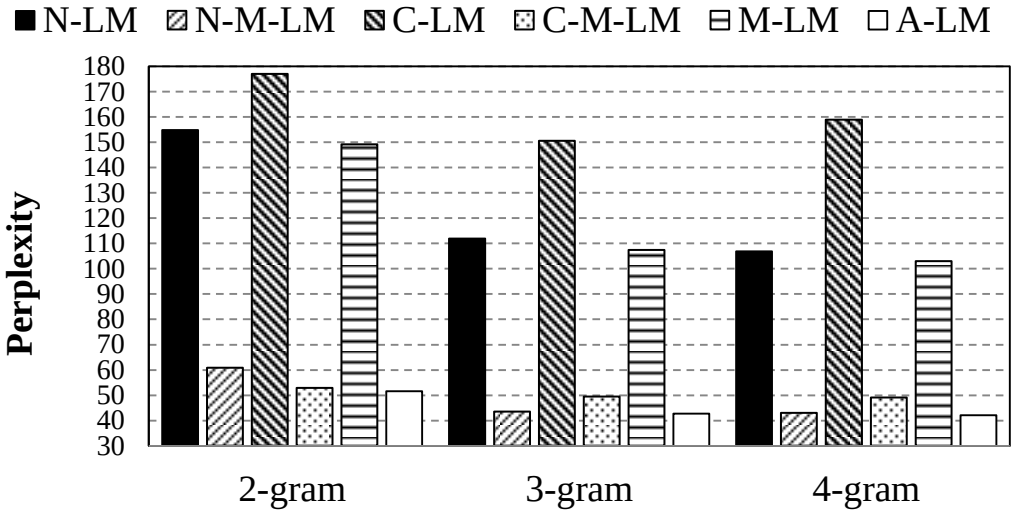

图 4. 6 YUCAI 测试集上混合语言模型的困惑度对比

从图 4.5 我们可以看到，在 XIXIANG 测试集上 A-LM 相对于 M-LM，困惑度降低了 47%；图 4.6 显示，在 YUCAI 测试集上 A-LM 相对于 M-LM，困惑度降低了 61.6%。

为了比较不同自适应语言模型在语音识别中的作用，我们从 XIXIANG 测试集中选取了对应的 15 位考生的真实音频，从 YUCAI 测试集中选取了对应的 11 位考生的真实音频用来作为语音识别的测试集。我们用单词错误率（Word Error Rate，WER）作为评价标准。特征提取采用的是 MFCC_E_D_A 的 39 维声学特征。声学模型包含了 16 个高

斯混合模型，共计 8000 个状态。语音识别实验在 HTK 下完成，进行说话人非监督性自适应，我们选取了三元文法的 M-LM 和 A-LM 作为对比，表 4.9 是不同测试上两个测试集的 WER 对比。表 4.9 说明，在 XIXIANG 测试集上，A-LM 的 WER 比 M-LM 降低了 16.9%；在 YUCAI 测试集上，A-LM 的 WER 比 M-LM 降低了 20.7%。这充分说明了本文提出的基于混合语言模型的语言模型自适应方法的有效性。

表 4. 9 M-LM 和 A-LM 在不同测试集上的 WER 对比

	M-LM	A-LM
XIXIANG	70.51%	53.57%
YUCAI	80.28%	59.59%

4.6 小结

本章详细的介绍了复述题型的特点及难点，复述题型自动评分流程的搭建过程并在复述题型数据库上完成了实验。针对复述题型自动评测中的两个难点，文本首先提出了内容相关的评分特征（包括 Similarity、KCR、WN 和 UWN），结合传统的基于发音质量和流利度评分特征集，在 RETELL-DB 数据库上取得了机器评分与人工评分为 0.621 的相关度，高于传统评分特征集的评分性能。随后，本文从语言学角度重点分析了考生在复述题型中的典型表现，针对复述题型自动评分过程中出现的非母语说话人自发性表述语音识别率不高的难点，从语言模型入手，提出了基于混合语言模型的语言模型自适应方法，并运用此方法构建了复述风格的语言模型。相比传统的语言模型建模方法，本文提出的方法最高将困惑度降低了 61.6%，并将语音识别的 WER 降低了 20.7%。

第五章 总结

本章首先简要介绍基于朗读题型发音自动评测技术和复述题型发音自动评测技术开发的艾尔斯口语考试系统，随后对论文工作进行总结，最后是对未来工作的展望。

5.1 艾尔斯口语考试系统

艾尔斯口语考试系统是基于英语发音自动评分系统开发的面向中学生群体，集制卷、考试、自动评分、反馈、检索于一体的英语口语考试软件系统。该软件分为客户端和服务器端，客户端采用 C#开发，基于.NET Framework 4.0，服务器端采用 Java 开发，搭建在 Linux 服务器分布式集群上。数据库采用的是 MySQL 数据库。

值得一提的是，本文还将艾尔斯口语考试系统推广至广东省深圳市育才中学，共计约 1500 名高中生在英语教学中使用了我们的系统，极大地推广了英语发音自动评测技术的在英语教学中的应用。图 5.1 是艾尔斯口语考试系统的系统功能流程图。

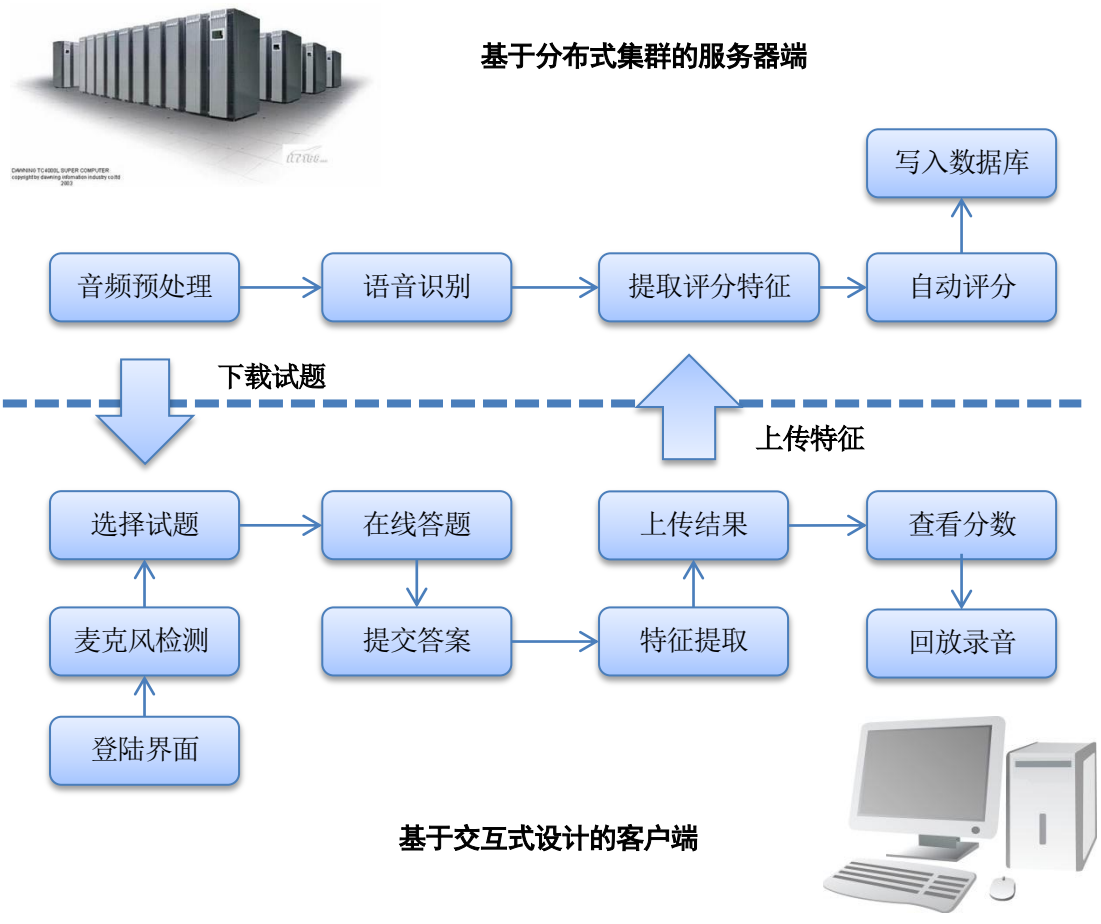

图 5.1 艾尔斯口语考试系统功能流程图

艾尔斯口语考试系统的软件截图如图 5.2~图 5.5 所示：

图 5.2 艾尔斯口语考试系统试题选择界面

图 5.3 艾尔斯口语考试系统朗读题界面

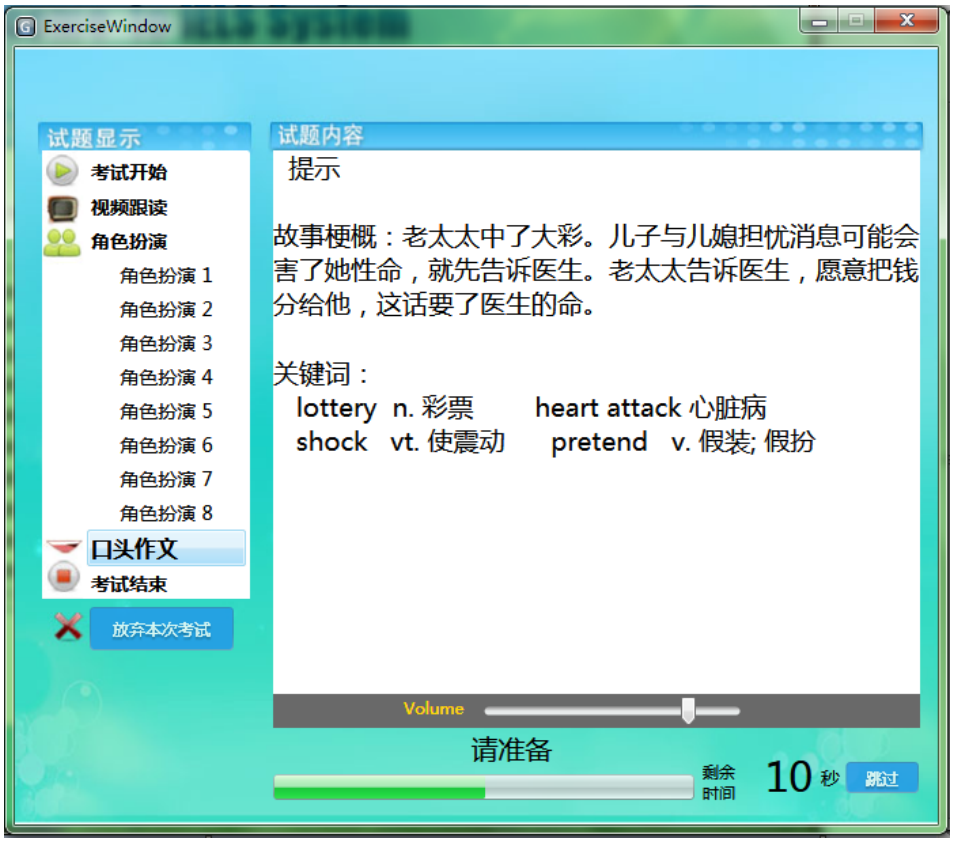

图 5.4 艾尔斯口语考试系统复述题界面

图 5.5 艾尔斯口语考试系统查看结果界面

5.2 论文工作总结

本文以朗读题型和复述题型为研究对象，研究了文本相关和文本无关的英语发音自动评测技术，搭建了朗读题型和复述题型的自动评测流程，并在实验数据库上取得了较好的评估结果。此外，基于英语发音自动评测系统，开发了艾尔斯口语考试系统，并推广到实际的英语教学中。至此，这里将本文的具体研究工作和研究成果归纳如下：

对于朗读题型自动评测问题，本文在主流技术的基础上，将研究问题集中在评分流程上，提出了基于强制对齐、标注语音段、非监督性说话人自适应、语音识别以及单音素循环网络解码操作的评分流程，并在此基础上提取了反映发音质量、流利度和内容相关的评分特征，最后利用多元线性回归拟合方法，得到机器评分结果。实验结果表明，本文的提出的方法在 READ-DB 语音数据库上获得了 0.923 的机器评分与人工评分的相关度，超过了传统基于强制对齐技术的自动评分方法，充分表明了本文所提出的评分流程的有效性。

对于复述题型的自动评测这一前沿研究领域，本文也做了较为深入的探索。首先，本文针对实际的数据库，深入的分析了复述题型自动评测的两大难点。针对提取高区分性的内容相关的评分特征的难点，本文提出了基于 Similarity、KCR、WN 和 UWN 四维内容相关的复述题型评分特征集，结合传统的基于发音质量和流利度的评分特征集，在 RETELL-DB 语音数据库上获得了 0.621 的机器评分与人工评分相关度，超过了传统评分特征集，这证明了本文提出的内容相关的评分特征集的有效性。对于提高复述题型中非母语说话人自发性表述的语音识别率的难点，本文将研究重点放在复述风格的语言模型构建方法上。首先从语言学角度分析了考生在复述题型中的典型表述特点，基于这些特点，本文选择和收集了话题相关文本语料、口语风格文本语料和书面风格文本语料，采用线性插值方法，提出了基于混合语言模型的语言模型自适应方法，并将方法应用于复述风格的语言模型构建中。实验结果表明，相比传统的语言模型建模方法，本文提出的方法最高将困惑度降低了 61.6%，并将复述题的语音识别的 WER 降低了 20.7%，充分证明了该方法的有效性。上述实验结果充分表明了本文在文本无关自动评测领域的探索研究具有非常积极的意义。

最后，基于英语发音自动评测方法的研究，本文开发了艾尔斯口语考试系统，并将系统推广至中学生日常的英语教学中，极大地推广了英语发音自动评测技术的应用，对中学生的英语教育事业有积极的意义。

5.3 对未来工作的展望

本文的工作包含了文本相关和文本无关的口语自动评测领域的研究内容，针对实际问题提出了较多的解决方法。然而，对于某些问题的处理还需要进一步的推敲。

对于复述题型自动评测问题，虽然本文针对两个难点问题都提出了一定解决方案，

但是可以看到，相比于朗读题型，复述题型的自动评测效果还不够如意，有很大的提升的空间。尤其是对于复述题型自动评测过程中需要解决的非母语说话人自发性表述的语音识别率问题，需要从语言模型和声学模型上做更多的改进工作。语言模型构建上，可以尝试基于人工引导词类聚类的词类语言模型，以及基于隐含语义分析的语言模型自适应方法；在声学模型的构建上，也需要充分考虑复述题型考生语音的发音特点，训练出更适合非母语说话人的声学模型。此外，在复述题型评分特征选取上，也可以尝试基于语义的评估方法，可以看到，复述题型的自动评测问题仍然是本领域最具有挑战性的研究工作，也是未来工作的研究重点。

参考文献

[1] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, “Automatic Pronunciation Scoring for Language Instruction”, in *Proc. ICASSP*, 1997, vol. 2, pp.1471-1474.

[2] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality", *Speech Communication*, Vol 30, Issues 2-3, 2000, pp.83-93.

[3] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, “Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech”, in *Proc. ICSLP*, 1996, pp. 1457-1460.

[4] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, “Combination of Machine Scores for Automatic Grading of Pronunciation Quality”, *Speech Communication*, 2000, Vol 30, Issues 2-3, pp. 121-130.

[5] C. Cucchiarini, H. Strik, and L. Boves, “Automatic Evaluation of Dutch Pronunciation by using Speech Recognition Technology”, in *Proc. ASRU*, 1997, pp. 622-629.

[6] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, “Automatic Evaluation and Training in English Pronunciation”, in *Proc. ICSLP*, 1990, pp. 1185–1188.

[7] J. Bernstein, J. DeJong, D. Pisoni, B. Townshend, “Two Experiments in Automatic Scoring of Spoken Language Proficiency”, In *Proc. InSTILL*, 2000, pp.57-61.

[8] K. Zechner, I. Bejar, “Towards Automatic Scoring of Non-Native Spontaneous Speech”, in *Proc. HLT-NAACL, 2006*.

[9] L. Chen, K. Zechner, and X. M. Xi, “Improved Pronunciation Features for Construct-Driven Assessment of Non-Native Spontaneous Speech”, in *Proc. HLT-NAACL, 2009*, pp. 442-449.

[10] K. Zechner, D. Higgins, X. M. Xi, and D. M. Williamson, ”Automatic Scoring of Non-Native Spontaneous Speech in Tests of Spoken English”, *Speech Communication, 2009,* vol. 51, issues 10, pp. 883-895.

[11] D. Higgins, X. M. Xi, K. Zechner, and D. M. Williamson, “A Three-stage Approach to the Automated Scoring of Spontaneous Spoken Responses”, *Computer Speech & Language*, vol. 25, issue 2, 2011, pp. 282-306.

[12] 严可,英文朗读题及复述题自动评测技术研究,硕士学位论文,中国科学技术大学, 2009.

[13] 丁克玉,李兆远,陈小平,胡国平,陈志刚,面向大规模英语口语考试的自动语法评分技术研究,*第十二届中国机器学习会议（CCML2010）*,2010.

[14] 江杰,口语测试自动评估技术研究,博士学位论文,中国科学院自动化所,2011.

[15] S. Xu, D. F. Ke, J. Jiang, X. Yang, H. Y. Li, and B. Xu, “Automatic Pronunciation Evaluation Based on Feature Extraction and Combination”, in *Proc. ICICIC*, 2008.

[16] J. Jiang and B. Xu, “Towards the Automatically Semantic Scoring in Language Proficiency Evaluation”, in *Proc. ICALT*, 2008, pp. 925-929.

[17] J. Mostow, *Project LISTEN(OL)*, 2008, from http://www.cs.cmu.edu/~mostow/.
[18] 安徽科大讯飞信息科技股份有限公司,畅言互动英语校园学习平台,见：http://www.iflytek.com/Html/cpfw/newyuyin/kypc/cyhdxx/.
[19] 宗成庆, 统计自然语言处理, 清华大学出版社, 北京, 2008.
[20] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book Version 3.4 Manual*, Cambridge, 2006.
[21] J. R. Bellegarda, "Statistical Language Model Adaptation: Review and Perspectives", *Speech Communication*, vol. 42, 2003, pp. 93-108.
[22] A. Park, T. Hazen, and J. Glass, "Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling", in *Proc. ICASSP*, 2005, vol. 1, pp.497-500.
[23] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals, "Transcription of Conference Room Meetings: An Investigation", in *Proc. INTERSPEECH*, 2005.
[24] T. Ng, M. Ostendorf, M. Y. Hwang, M. Siu, I. Bulyko, and X. Lei, "Web-Data Augmented Language Models for Mandarin Conversational Speech Recognition", in *Proc. ICASSP*, 2005, pp. 589-592.
[25] G. Moore and S. Young, "Class-based Language Model Adaptation using Mixtures of word-class Weights", in *Proc. ICSLP*, 2000, pp. 512-515.
[26] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram Models of Natural Language", *Computational Linguistics*, vol. 18, no. 4, 1992, pp.467-479.
[27] Y. Akita and T. Kawahara, "Efficient Estimation of Language Model Statistics of Spontaneous Speech via Statistical Transformation Model", in *Proc. ICASSP*, Toulouse, France, 2006.
[28] H. Schramm, X. L. Aubert, C. Meyer, and J. Peters, "Filled-Pause Modeling for Medical Transcriptions", in *Proc. SSPR*, 2003, pp. 143-146.
[29] K. Ohta, M. Tsuchiya, and S. Nakagawa, "Construction of Spoken Language Model Including Fillers Using Filler Prediction Model", in *Proc. INTERSPEECH*, 2007, pp.1489-1492.
[30] 杨军, 口语非流利产出研究评述, 外语教学与研究, 第 36 卷, 第 4 期, 2004.
[31] 戴朝晖, 中国大学生汉英口译非流利现象研究, 上海翻译, 第 1 期, 2011.
[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*. NTIS order number PB91-100354, 1993.
[33] D. Paul and J. Baker, "The Design for the Wall Street Journal-based CSR Corpus", *DARPA Speech & Nat. Lang. Workshop*, Arden House, NY, 1992.
[34] S. Witt and S. Young, "Computer-assisted Pronunciation Teaching based on Automatic Speech Recognition", in *Language Teaching and Language Technology*, 1997, pp. 25-35.

[35] L. Wasserman, *All of Statistics*, Springer-Verlag New York, Inc, 2004.
[36] The R Project for Statistical Computing, http://www.r-project.org/.
[37] Q. F. Wen, M. C. Liang, and X. Q. Yan, *Spoken and Written English Corpus of Chinese Learners (SWECCL) 2.0*, Foreign Language Teaching and Research Press, Beijing, China, 2008.
[38] A. Stolcke, "SRILM – an extensible language modeling toolkit", in *Proc. ICSLP*, Denver, 2002, pp. 901-904.
[39] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text Compression*. Prentice Hall, Englewood Cliffs, N. J.
[40] R. Kneser and H. Ney, "Improved Backing-off for m-gram Language Modeling", in *Proc. ICASSP*, 1995, pp.181-184.

作者简历及攻读学位期间发表的学术论文与研究成果

姓名：陈蒙　　性别：男　　出生日期：1986.12.20　　籍贯：湖南道县

2009.9 -- 2012.7	中科院深圳先进技术研究院	计算机应用技术	工学硕士
2005.9 -- 2009.7	北京科技大学	计算机科学与技术	工学学士

【攻读硕士学位期间发表的论文】

[1] Meng Chen, Yang Song and Lan Wang, "Adapted Language Modeling for Recognition of Retelling Story in Language Learning", in *Proc. 3rd International Conference on Audio, Language and Image Processing (ICALIP 2012)*, 2012. (Submitted)

[2] Meng Chen, Dean Luo and Lan Wang, "Automatic Scoring in a Task of Retelling Stories for Language Learners," in *Proc. 15th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2011)*, 2011.

[3] 陈蒙，罗德安，王岚，英语口语考试中复述题自动评测技术的研究，*先进技术研究通报*，第五卷，第 10 期，2011 年；

[4] 魏三喜，孙剑，吴海飞，刘振智，陈蒙，罗德安，王岚，基于自动评分的交互式英语学习系统，*先进技术研究通报*，第五卷，第 10 期，2011 年。

【攻读硕士学位期间参加的科研项目】

英语口语自动判分系统（企业合作项目），2010 年 9 月~2011 年 9 月

【攻读硕士学位期间的获奖情况】

[1] 2012 年获得中国科学院深圳先进技术研究院“院长奖学金优秀奖”

[2] 2011 年被评为中国科学院深圳先进技术研究院“优秀三好学生”

【软件著作权】

王岚，陈蒙，金晓虎，艾尔斯口语考试系统，计算机软件著作权，2012。（已提交）