基于生成式人工智能的大学英语四、六级听力测试开发及资源库建设研究

过君琰

（无锡学院，江苏 无锡 214105）

摘 要: 生成式人工智能正以前所未有的广度和深度重塑外语教育的生态，为外语教学注入全新的活力与可能性。然而在外语测试领域，其应用与实践尚处于探索阶段。本研究探讨了运用生成式人工智能（ChatGPT、DeepSeek 和 Murf AI）开发符合内容效度要求的大学英语四、六级听力测试的可行性。结果表明，ChatGPT 和 DeepSeek 能够生成符合四、六级听力任务特征的文本与试题，Murf AI 合成的音频在口音、语速和清晰度方面基本满足测试需求。基于研究结果，可构建听力测试资源库，为听力测试开发提供可持续更新的资源支持。

关键词：生成式人工智能；大学英语考试；听力测试；资源库

基金项目：2024 基于人工智能的大学外语数字化创新教学专项课题(2024RGWY045)

DOI：doi.org/10.70693/jyxb.v2i2.369

Research on the Development of CET-4/6 Listening Tests and the Construction of Resource Banks Based on Generative Artificial Intelligence

Junyan Guo

Department of General Education, Wuxi University, Wuxi, China

Abstract: Generative Artificial Intelligence (GenAI) is reshaping the ecology of foreign language education with unprecedented width and depth, injecting new vitality and possibility into foreign language teaching and learning. However, in the field of foreign language testing, its application remains in the exploratory stage. This study investigates the feasibility of applying GenAI (ChatGPT, DeepSeek and Murf AI) to develop content-valid listening tests of the College English Test (CET). The results showed that ChatGPT and DeepSeek can generate scripts and test items aligned with the task characteristics of CET-4 and CET-6 listening assessments. The audio files synthesized by Murf AI met the testing requirements in terms of accent, speech rate and clarity. Based on the research findings, a listening test resource bank can be established to provide sustainably updated resources for listening test development.

Keywords: Generative Artificial Intelligence (GenAI); College English Test (CET); Listening test; Resource bank

作者简介：过君琰(1984—)，女，博士，研究方向外语教学、外语测试。

通讯作者：过君琰

一、引言

生成式人工智能的发展给外语教育带来了全面、深刻和系统的变化与契机。生成式人工智能是指一组人工智能算法和模型，它们具备类似人类的创造力与适应性，能够生成新内容，包括文本、图、视频以及问题解决策略等[1]。从理论到实践层面，外语界掀起了一股关于生成式人工智能在外语教育应用的探讨热潮。在理论层面，学者们从语言本体特征[2]、二语习得机制[3]、外语教学实践[4][5][6]、理论框架构建[7][8]以及对人与科技关系的反思[9]等不同方面切入，深入剖析生成式人工智能与外语教育融合的内在机理，挖掘其深度价值。在实践领域，学者们探索了生成式人工智能在认知与态度[10][11]、语言技能培养[12]、教学资源开发[13][14]、教学实施与融合[15]、教师发展[16]等方面，为生成式人工智能融入外语教育的创新发展提供了新思路[17]。然而，与生成式人工智能在外语教学领域的快速发展相比，其在外语测试领域的系统应用相对滞后，虽然关于各项技能的研究已渐次展开，例如翻译[18]、阅读[19]和口语[20]等，但主要集中于写作，例如协助写作过程[21]、实施自动评分[22]、提供写作反馈[23]等，听力测试领域相关研究相对较为匮乏[24]，尚未形成系统成果。鉴于此，本研究探讨如何基于大学英语四、六级真题语料库特征分析，运用生成式人工智能开发符合测评内容效度要求的大学英语四、六级听力测试，并在此基础上构建可扩展的测试资源库，以期提升测试命题效率，为听力试题开发提供建议和参考。

二、文献综述

对真题语料库的任务特征分析能够让我们理解语言测试的命题规律和语言特征，确保测试的内容效度，而生成式人工智能凭借其在自然语言理解、分析、模拟和生成的强大能力，为听力测试的材料生成和题目设计提供了新的思路，推动听力测试向智能化方向发展。

（一）听力真题语料库的特征分析

听力测试语料库是指系统编制的口语语料电子数据库。基于真实考试材料的语料库能够较为全面地反映测试所取样的语言能力范围，为检验测试效度、信度与测试目标的一致性提供实证依据。通过对真题语料的系统标注与统计分析，研究者可以了解试题在话题、语篇及交际功能等方面的任务特征，为后续测试的开发提供理论支撑与实证依据。关于听力真题语料库特征分析的既有研究大体可以归纳为两个研究方向。

首先是研究者们构建由高风险测试多年真题组成的语料库，这些测试包括高考、大学英语四、六级和雅思考试等，然后分析其中的语言模式和文本特征为语言测试与评估提供依据。辜向东、李亚果的两项研究围绕 2005 年的大学英语四、六级改革，分析了 1996 年至 2008 年四、六级听力真题的语篇输入和预期回答两方面的任务特征[25][26]。过君琰对 2016 年至 2024 年的四、六级听力真题进行上述两方面的分析，其中语篇输入包括体裁、主题、词汇、难度、话轮数和语速，预期回答包括考查技能[27]。Tao 和 Aryadoust 分析了从 2000 年至 2022 年 31 省（自治区、直辖市）的 170 套高考听力材料的语言特征及其对应的功能维度[28]。Aryadoust 通过分析 1996 至 2021 年 256 份雅思听力考试的主题和口音，为提高测试的公平性和公正性提出建议[29]。

另外一个研究方向是测试开发者和研究者运用具有代表性的语料库作为高风险听力测试的参照，以衡量这些测试的内容效度。内容效度是指考试内容是否涵盖了课程教学要求或考试大纲所规定的语言技能、语言结构等内容[30]。这些语料库包括英国国家语料库(British National Corpus, BNC)、当代美国英语语料库(Corpus of Contemporary American English, COCA)、国际学习者英语语料库(International Corpus of Learner English, ICLE)和密歇根大学学术口语语料库(Michigan Corpus of Academic Spoken English, MICASE)。例如 Nguyen 的研究比较了密歇根大学学术口语语料库和雅思考试听力第三、第四部分的词块[31]，而 Zhao 和 Aryadoust 的研究则基于自动化语义分析，考察该语料库和雅思以及托福听力部分微型讲座的语义特征[32]。

（二）生成式人工智能与听力测试

目前生成式人工智能在听力测试领域的应用尚处于起步阶段。已有的少数研究表明，其在自动生成多样化听力测试文本、题目以及选项方面展现出一定的潜力。许家金等总结了大语言模型在听力教学中的应用，并以 ChatGPT 4 为例探讨如何改编材料、设计习题、制作音频和词表[33]。Aryadoust 等讨论了 ChatGPT 4 在为不同水平的受试者（学术型、低水平、中等水平和高水平）生成听力文本和测试试题方面的能力[34]。同时研究

者们对 ChatGPT 4 产生的文本就语言学特征、主题变异性和选项重叠程度方面进行了分析。结果表明，ChatGPT 4 能够较为稳定地生成针对不同水平的具有显著语言差异的文本，但在不同选项之间存在语义重叠的现象，因此四个难度水平在题项层面和文本层面均未呈现显著差异。Runge 等运用 ChatGPT 3 生成多邻国英语测试 (Duolingo English Test, DET) 中交互式听力测试内容[35]。这种基于情景的对话任务包含不同的对话者（学生对学生、学生对教授）。在每一轮的对话中，受试者需要从一些选项中选择合适的使对话继续。在任务的最后，受试者对对话进行进行总结。这项大规模的试点研究包含 713 个测试任务，每个任务收集了 464 份作答。研究结果表明采用人机协同的方法，能够在大规模条件下自动化生成测试内容，从而为大语言模型自动化生成题项的可行性提供了实证依据。Wang 和 Meng 的研究则探讨通过生成式人工智能（Kimi）和人工专家相结合是否能够提高听力选择题干扰项质量[36]。研究结果表明虽然生成式人工智能在充分捕捉听力误解模式以及情境化语言使用方面存在不足，但它能够保持内容与结构的一致性并确保语义上的相对独立性。因此在基于原则的提示语和人类监督相结合的前提下，生成式人工智能能够提升干扰项的质量。

基于前述研究综述，本研究以大学英语四、六级听力真题为研究对象，在笔者前期系统分析其任务特征成果的基础上[27]，探讨如何运用生成式人工智能工具开发听力测试，以期为听力测评的设计提供新的思路与方法。

三、生成式人工智能辅助四、六级听力测试开发

《全国大学英语四、六级考试大纲（2016 年修订版）》（以下简称《考纲》）规定大学英语四、六级听力考试各包含三个部分，四级为短篇新闻、长对话和听力篇章，六级为长对话、听力篇章和讲话/报道/讲座[37]。测试题型均为选择题（单选题）。本节以两个级别考试中共有的题型—长对话和听力篇章为例，探讨如何运用 ChatGPT 和 DeepSeek 进行听力文本生成和题目设计以及运用 Murf AI 进行音频录制。

（一）文本生成

1. 长对话

《考纲》规定长对话每篇的词数分别为四级 240 至 280 词和六级 280 至 320 词。大学英语四、六级对标欧洲语言共同参考框架（The Common European Framework of Reference for Languages, CEFR）分别为 B1 和 B2 水平[38]。四级长对话的听力真题参考文本为 2024 年 6 月第 2 套第 1 个对话，该对话的主题为教导孩子如何存钱和花钱。提示语设置为 Please generate a listening script of a conversation on the topic of “teach children…” with a length of 240-280 words, designed for learners at the CEFR B1 level. A sample script is provided below as a reference。六级长对话听力真题参考文本为 2023 年 6 月第 2 套第 1 个对话，该对话的主题为教授与学生之间的谈话：选择建筑学专业及其准备工作。提示语为 Please generate a listening script of a conversation between a professor and a student with a length of 280-320 words, designed for learners at the CEFR B2 level. A sample script is provided below as a reference。

根据过君琰的分析框架[27]，笔者从主题、词汇、难度和话轮数四个维度对比分析真题长对话和 ChatGPT、DeepSeek 各自生成长对话的内容效度，从主题、词汇和难度三个维度分析真题听力篇章和生成式人工智能生成的篇章。其中词汇的形符数、类符数和词频覆盖率通过 Vocabprofilers (https://www.lextutor.ca/vp/comp/) 的 BNC-COCA (1-25k)语料库进行统计，Nation 认为 98%以上的词汇覆盖率是理想的[39]。文本难度使用英文分级指难针统计[40]。指难针运用阅读真题文本进行词汇、句法和篇章的核心特征进行难度计算，确定难度级别范围，并将其与《中国英语能力等级量表》的三至七级衔接[41]。数值 3.00-3.99 之间表示三级（中考）难度，4.00-4.99 是四级（高考），5.00-5.99 是五级（CET-4），6.00-6.99 是六级（CET-6），7.00-7.99 是七级（考研）。需要注意的是生成式人工智能生成的文本偶尔会不在提示语所规定的词数范围内，因此笔者会再一次给出词数范围的指令，要求重新生成。文本难度如果过高则可以根据英文分级指难针给出的难词和难句简化建议进行调整。如果难度过低则可以再给提示语，让其调整各自生成文本的难度。表 1 和表 2 的结果显示，经过调整，ChatGPT 和 DeepSeek 能够生成和真题长对话的任务特征基本一致的文本。

表 1 ChatGPT 和 DeepSeek
生成四级长对话任务特征对比

任务特征	真题	ChatGPT	DeepSeek
主题	教导孩子	教导孩子养	教导孩子

		如何存钱和花钱	成健康的饮食习惯	注意安全
词汇	形符数	283	278	276
	类符数	165	163	162
	词频覆盖率	最高频 3,000 词覆盖 98.2%	最高频 3,000 词覆盖 98.6%	最高频 3,000 词覆盖 98.2%
难度		4.94	4.74	4.96
话轮数		5.5	5.5	5.5

表 2 ChatGPT 和 DeepSeek
生成六级长对话任务特征对比

任务特征		真题	ChatGPT	DeepSeek
主题		教授与学生对话:选择建筑学专业及其准备工作	教授与学生对话：学术研究选题与研究方案设计	教授给学生提供论文反馈
词汇	形符数	312	304	316
	类符数	172	196	180
	词频覆盖率	最高频 3,000 词覆盖 98.1%	最高频 3,000 词覆盖 99%	最高频 3,000 词覆盖 99.4%
难度		4.79	4.96	4.84
话轮数		5.5	5.5	5.5

2. 听力篇章

根据《考纲》听力篇章每篇的词数分别为四级 220 至 240 词和六级 240 至 260 词。按照长对话的模式，笔者选取四、六级真题听力篇章各 1 篇作为参考文本，分别运用 ChatGPT 和 DeepSeek 生成相应的听力篇章。四级的参考文本为 2023 年 12 月第 2 套第 3 篇，该篇章的主题为个人空间。提示语设置为 Please generate a listening script of a passage with a length of 220-240 words, designed for learners at the CEFR B1 level. A sample script is provided below as a reference。六级为 2023 年 12 月第 1 套第 1 篇，该对话的主题为作家凯特•阿特金森(Kate Atkinson)。提示语为 Please generate a listening script of a passage with a length of 240-260 words, designed for learners at the CEFR B2 level. A sample script is provided below as a reference。表 3 和表 4 的结果显示，ChatGPT 和 DeepSeek 能够生成和真题听力篇章的任务特征基本一致的文本。

表 3 ChatGPT 和 DeepSeek
生成四级听力篇章任务特征对比

任务特征		真题	ChatGPT	DeepSeek
主题		个人空间	身势语	睡眠现象—说梦话
词汇	形符数	240	228	231
	类符数	134	145	158
	词频覆盖率	最高频 4,000 词覆盖 98.8%	最高频 4,000 词覆盖 98.7%	最高频 4,000 词覆盖 99.6%
难度		5.09	5	5.01

表 4 ChatGPT 和 DeepSeek
生成六级听力篇章任务特征对比

任务特征		真题	ChatGPT	DeepSeek
主题		作家凯特•阿特金森	作家奇玛曼达•恩戈齐•阿迪奇埃	作家石黑一雄
词汇	形符数	257	254	260
	类符数	159	180	179
	词频覆盖率	最高频 5,000 词覆盖 98.1%	最高频 5,000 词覆盖 98%	最高频 5,000 词覆盖 98.1%
难度		6.7	6.71	6.85

（二）题目设计

1. 理解明示的信息

《考纲》中提到听力考核技能包括理解明示的信息、理解隐含的信息、运用语言特征理解听力材料和运用听力策略。根据过君琰的研究结果[27]，四、六级真题题目的考核技能主要集中在理解明示的信息（超过 97%）和理解隐含的信息（2% 左右），而理解明示的信息中又以听懂重要信息或特定细节为主。本节主要以听力篇章展示运用 ChatGPT 和 DeepSeek 生成题目，考核理解明示的信息（听懂重要信息或特定细节）以及理解隐含的信息两种技能。如例 1 所示，表 3 中四级真题听力篇章的第 2 个问题是关于一个只是泛泛之交的人进入我们的亲密空间的细节，答案是 A. It makes us feel uncomfortable, 与听力文本中的信息一致（A mere acquaintance entering our intimate space makes us uncomfortable）。

例 1： What happens if a mere acquaintance enters our intimate space?

A. It makes us feel uncomfortable.

B. It renders the acquaintance a stranger.

C. It brings the acquaintance closer to us.

D. It causes the bubble around us to vanish.

笔者运用 ChatGPT 和 DeepSeek 分别对表 3 中其各自生成的四级听力文本提出一个关于重要信息或特定细节的问题。提示语是：Please create a multiple-choice question focusing on a specific detail from the following passage, with four options。例 2 是 ChatGPT 给出的问题以及选项，答案是 C. Defensiveness or discomfort，与听力文本中的细节 while crossed arms may signal defensiveness or discomfort 一致。例 3 是 DeepSeek 给出的问题以及选项，答案是 B. During periods of stress, illness, or significant lack of sleep，与听力文本中给出的信息一致(They may become more frequent during periods of stress, illness, or significant lack of sleep)。

例 2： According to the passage, what might crossed arms suggest?

A. Confidence and self-assurance.

B. Agreement and support.

C. Defensiveness or discomfort.

D. Happiness and excitement.

例 3： According to the passage, when do episodes of sleep talking tend to become more frequent?

A. During periods of dreaming.

B. During periods of stress, illness, or significant lack of sleep.

C. After intense physical exercise.

D. When trying to remember dreams.

2. 理解隐含的信息

笔者运用 ChatGPT 和 DeepSeek 分别对表 4 中其各自生成的六级听力文本提出一个关于推论隐含意义的问题，提示语是 Please create a multiple-choice question about an implied meaning or inference from the following passage, with four options。例 4 是 ChatGPT 给出的问题及选项。正确答案是 B。文本中的信息(As a teenager, she read extensively, particularly British and American novels, before gradually discovering African writers whose works connected more closely with her own cultural experience.)并没有直接说明发现非洲作家产生重大影响，听者根据上述信息需要推断出与文中作者此前阅读的英美小说相比，非洲作家的作品对她来说更具个人意义。

例 4： Based on the passage, why did discovering African writers have a significant impact on Adichie as a teenager?

A. Their works were easier to understand than British and American novels.

B. Their stories reflected experiences and cultural backgrounds closer to her own.

C. They focused mainly on political science, which she later studied.

D. They encouraged her to move to the United States for further education.

例 5 是 DeepSeek 给出的问题及选项，正确答案是 B。根据文本中的信息(contemplation of morality, profound inquiries into the nature of forgetting, his novels share a fundamental emotional depth)可以推断无论作家石黑一雄用什么题材（科幻、奇幻、历史），他最终关注的始终是人类共同的情感深度和人生课题。

例 5： According to the passage, why does the author suggest Kazuo Ishiguro uses fantasy or science fiction settings in his later novels?

A. To reach a wider international audience.

B. To explore deep and universal human themes.

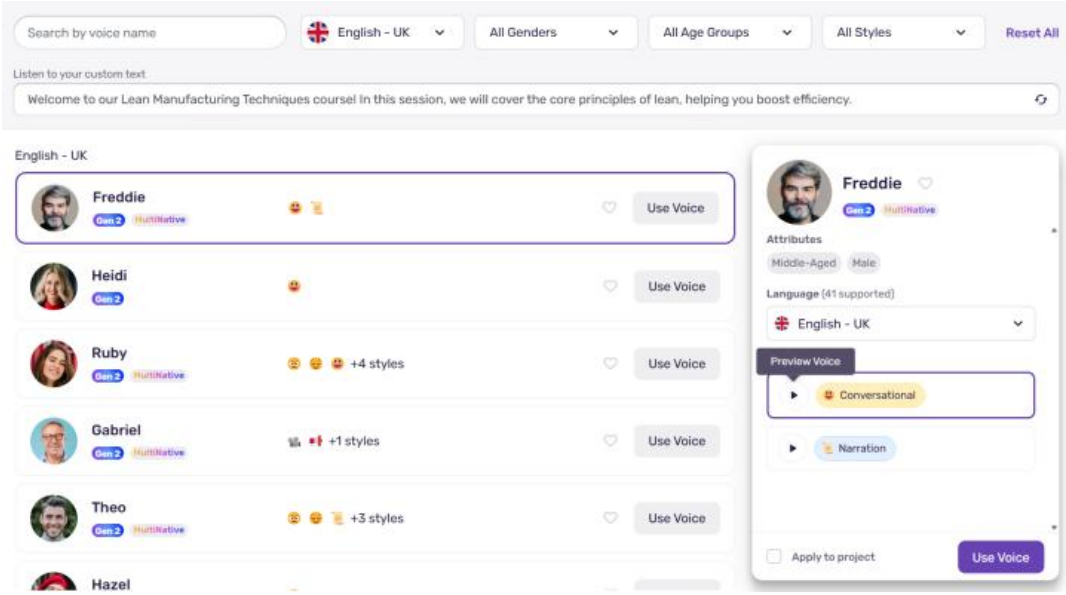
C. To move away from his Japanese background.

D. To make his stories more entertaining.

（三）音频录制

本节展示如何运用 Murf AI 将 ChatGPT 和 DeepSeek 生成的文字材料转换为音频格式。Murf AI 是一个基于云的平台，提供先进的文本转语音技术，利用人工智能的机器学习生成逼真自然的 AI 配音。它提供超过 300 种 AI 语音，涵盖超过 33 种语言，声音比较清晰自然。在 Murf AI 中需要先创建项目，然后选择朗读声音。由于《考纲》规定四、六级录音材料使用标准的英式或美式英语朗读，因此我们在标签栏选择 English－US & Canada 或者 English－UK。在标签栏我们还可以选择朗读者的性别（如：男、女）、年龄（如：年轻、中年）以及声音风格（对话式等 17 种）。如图 1 所示，Freddie 是一位中年男性，我们可以在右下角的风格选项中预听他不同风格的声音(Preview Voice)。对于新闻听力建议选择 Newscast 风格，长对话更适合 Conversational，听力篇章则可以选择 Narration。

图 1 Murf AI 声音选择界面



确定使用该声音之后就可以把生成的文本复制到文本框里，点击右侧的小三角就可以制作音频了。根据《考纲》规定，四级听力语速约为每分钟 120-140 词,每个问题后留有 15 秒答题时间。六级语速为每分钟 140-160 词，每个问题后留有 13 秒答题时间。如图 2 所示，所选声音旁边的标签栏可以调整音高(Pitch)、语速(Speed)和增加停顿时间(Add Pause)。新闻和听力篇章文本可以一句、一段或整篇录制。长对话需要根据两位对话者的内容分层交替将文本粘贴在一个个方框(Block)里面（图 3）。录音完成之后就可以下载导出了。

图 2 Murf AI 音频录制界面（新闻、听力篇章）

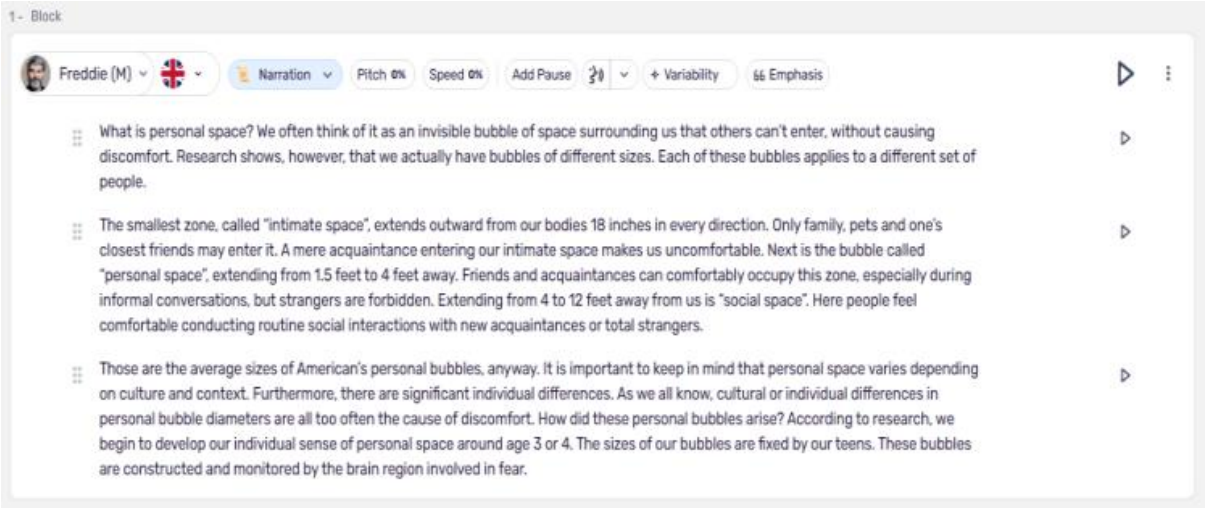
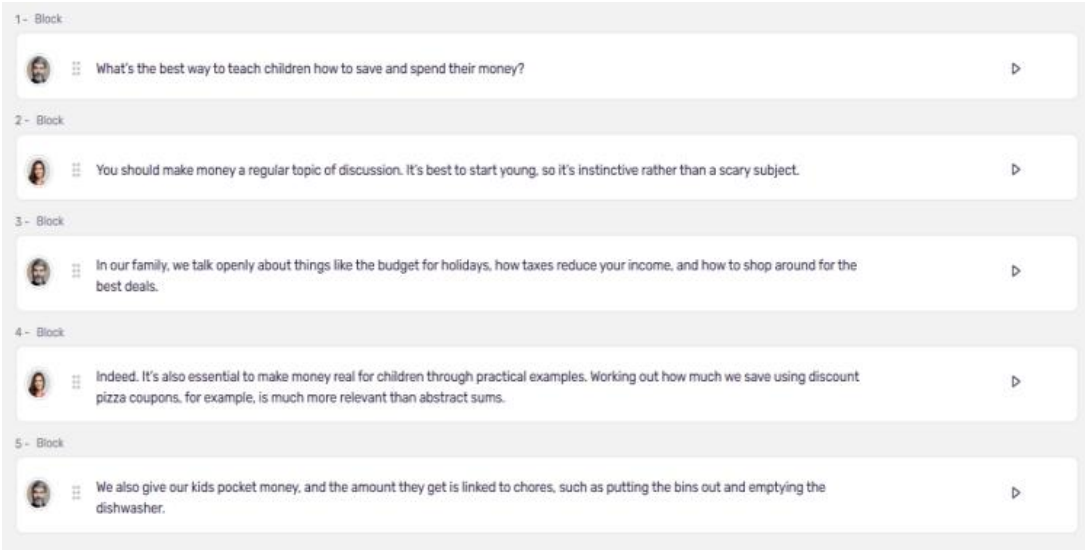


图 3 Murf AI 音频录制界面（长对话）



四、结论

本文探讨了如何运用生成式人工智能开发符合测评内容效度要求的大学英语四、六级听力测试。具体而言，本研究运用 ChatGPT 和 DeepSeek 生成听力文本并设计相关试题，同时利用 Murf AI 将文本转为语音，实现听力材料的音频化制作。研究结果发现，生成式人工智能在听力文本生成与试题设计方面具有较高的效率和灵活性，经过调试能够产出符合大学英语四、六级听力任务特征的文本和考核技能的题目。同时，使用 Murf AI 进行语音合成能够有效提高音频制作效率，其口音、语速和语音清晰度方面基本满足测试需求。

尽管本研究显示生成式人工智能在大学英语四、六级听力测试开发中展现了一定的应用潜力，但仍存在一定的局限性。首先，生成式人工智能输出内容对提示语高度敏感，不同指令设定可能导致文本难度、语篇结构及题目质量产生差异，因此在不同内容主题与任务情境下生成结果的稳定性与可重复性仍有待进一步检验。尽管 ChatGPT 和 DeepSeek 能够较好地模拟听力语篇特征，但它们能否把握任务的语用特征、真实性和文化语境等层面还需要进一步探索。其次，本研究尚未开展大规模考生数据收集与统计分析，因此缺乏基于统计数据的量化证据支撑，难以评估基于生成式人工智能开发的听力测试在预测或区分考生真实语言能力方面的效力。

针对上述局限，未来研究可以从以下几个方面拓展。首先，进一步优化生成式人工智能的提示语设计和生成流程，提高输出文本在语篇结构和题目设计方面的一致性和可控性。同时可以进一步比较不同模型在生成文本和题目方面的差异。其次，应该在不同语言水平的考生群体中开展实证研究，对生成式人工智能生成的听力材料进行项目分析与信度检验。同时对比 AI 录音与真实录音在语调、连贯性以及自然度上的差异，评估其对听力理解和考生表现的影响。再次，可以通过问卷和访谈研究受试者和测试专家对生成式人工智能生成的试题和录音在真实性和可信度方面的感知。借助生成式人工智能根据不同主题生成与大学英语四、六级听力考试匹配的文本、试题和音频材料经过系统化质量评估与试测，可进一步进行整理、分类和整合入库，逐步构建一个主题多样、难度分级、结构规范的听力测试资源库，为后续听力测试开发提供标准化、可持续的素材支持。

参考文献：

[1] HE R, CAO J, TAN T. Generative artificial intelligence: A historical perspective [J]. National Science Review, 2025, 12(5), nwaf050. https://doi.org/10.1093/nsr/nwaf050

[2] 冯志伟, 张灯柯. GPT 与语言研究[J]. 外语电化教学, 2023, (2): 3-11+105. https://doi.org/10.20139/j.issn.1001-5795.2023.02.001

[3] 杨连瑞. ChatGPT 大语言模型背景下的二语习得[J]. 现代外语, 2024, 47(4): 578-585. https://doi.org/10.20071/j.cnki.xdwy.20240523.003

[4] 张震宇, 洪化清. ChatGPT 支持的外语教学：赋能、问题与策略[J]. 外语界, 2023, (2): 38-44. https://doi.org/10.26971/j.cnki.flw.1004-5112.2023.02.006

[5] 文秋芳. 人工智能时代的英语教育：四要素新课程模式解析[J]. 中国外语, 2024, 21(3): 1, 11-18.

[6] 胡壮麟. ChatGPT 谈外语教学[J]. 中国外语, 2023, (3): 1+12-15. https://doi.org/10.13564/j.cnki.issn.1672-9382.2023.03.003

[7] 李炜炜. 人工智能赋能外语教育改革：理念创新与行动逻辑[J]. 中国高等教育, 2023, (9): 49－52.

[8] 郑咏滟. 生成式人工智能在外语教育中的应用：关键争议与理论构建[J]. 外语教学, 2024, 45(6): 48-53. https://doi.org/10.16362/j.cnki.cn61-1023/h.2024.06.012

[9] THORNE S L. Generative artificial intelligence, co-evolution, and language education[J]. The Modern Language Journal, 2024, 108(2): 567-572. https://doi.org/10.1111/modl.12932

[10] PRILOP C N, MAH D-K, JACOBSEN L J, et al. Generative AI in teacher education: Educators' perceptions of transformative potentials and the triadic nature of AI literacy explored through AI-enhanced methods[J]. Computers and Education: Artificial Intelligence, 2025, 9, 100471. https://doi.org/10.1016/j.caeai.2025.100471

[11] 张蕊, 刘雅楠. ChatGPT 辅助大学生英语写作中心流体验、自我效能感与学业浮力的作用机制研究[J]. 西安外国语大学学报, 2025, 33(3): 61-67. https://doi.org/10.16362/j.cnki.cn61-1457/h.2025.03.005.

[12] LI S. Generative AI and second language writing[J]. Digital Studies in Language Literature, 2025, 2(1): 122-152. https://doi.org/10.1515/dsll-2025-0007

[13] LO A W T. The educational affordances and challenges of generative AI in Global Englishes-oriented materials development and implementation: A critical ecological perspective[J]. System, 2025, 130, 103610. https://doi.org/10.1016/j.system.2025.103610

[14] 贾蕾, 马颖. 生成式人工智能在外语教材编写中的应用[J]. 外语研究, 2025, 42(2): 55-61+113. https://doi.org/10.13978/j.cnki.wyyj.2025.02.008

[15] 于晖, 宋金戈. 知识共建视域下生成式人工智能辅助的外语教学—定位、应用与发展[J]. 中国外语, 2025, 22(3): 4-14. https://doi.org/10.13564/j.cnki.issn.1672-9382.2025.03.004

[16] MOORHOUSE B L, WAN Y, WU C, et al. Developing language teachers' professional generative AI competence: An intervention study in an initial language teacher education course[J]. System, 2024, 125, 103399. https://doi.org/10.1016/j.system.2024.103399

[17] LI B, TAN Y L, WANG C, et al. Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching[J]. Computers and Education: Artificial Intelligence, 2025, 9, 100445. https://doi.org/10.1016/j.caeai.2025.100445

[18] 王勇, 王雨晨. 生成式 AI 在大学英语翻译教学中的应用研究[J]. 长春工程学院学报(社会科学版), 2025: 26(2), 122-128.

[19] LIN Z, CHEN H. Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items[J]. System, 2024, 123, 103344. https://doi.org/10.1016/j.system.2024.103344

[20] ARYADOUST V, ZAKARIA A. ChatGPT in the assessment of speaking[C]. MCCALLUM L, TAFAZOLI D. (Eds.), The Palgrave Encyclopedia of computer-assisted language learning, 2025: 1-8. Switzerland: Springer Nature. https://doi.org/10.1007/978-3-031-51447-0_200-1

[21] CONNEL PENSKY A E, USDAN J H, CHANG H. Generative AI's impact on graduate student professional writing productivity and quality[J]. International Journal of Artificial Intelligence in Education, 2025, 35(6): 4057-4082. https://doi.org/10.1007/s40593-025-00528-z

[22] BARROT J S. Generative artificial intelligence for automated essay scoring: Exploring teacher agency through an ecological perspective[J]. Assessing Writing, 2026, 67, 100990. https://doi.org/10.1016/j.asw.2025.100990

[23] CROSTHWAITE P, SUN S. Generative AI and L2 written feedback studies: A scoping review[J]. RELC Journal, 2025, 00336882251386530. https://doi.org/10.1177/00336882251386530

0

[24] GOH C C M, ARYADOUST V. Developing and assessing second language listening and speaking: Does AI make it better?[J] Annual Review of Applied Linguistics, 2025, 45: 179-199. https://doi.org/10.1017/S0267190525100111

[25] 辜向东, 李亚果. 改革后 CET 听力测试语篇输入与预期回答任务特征分析[J]. 西安外国语大学学报, 2010, 18(4): 71-74+79. https://doi.org/10.16362/j.cnki.cn61-1457/h.2010.04.024

[26] 辜向东, 李亚果. CET 听力测试语篇输入和预期回答任务特征历时分析(1996-2007)[J]. 外语测试与教学, 2012, (3): 17-26. https://doi.org/10.26970/j.cnki.fltt.1004-5112.2012.03.004

[27] 过君琰. 大学英语四、六级听力测试命题质量历时研究[J]. 人文与社会科学学刊, 2026, 2(2): 32-38. http://doi.org/10.70693/rwsk.v2i2.202

[28] TAO X, ARYADOUST V. A multidimensional analysis of a high-stakes English listening test: A corpus-based approach. Education Sciences[J], 2024, 14(2): 137. https://doi.org/10.3390/educsci14020137

[29] ARYADOUST V. Topic and accent coverage in a commercialized L2 listening test: Implications for test-takers' identity[J]. Applied Linguistics, 2024, 45(5): 765-785. https://doi.org/10.1093/applin/amad062

[30] HUGHES A, HUGHES J. Testing for language teachers (3rd ed.)[M]. Cambridge: Cambridge University Press, 2020: 29-30. https://doi.org/10.1017/9781009024723

[31] NGUYEN P H T. Investigating the content validity of the IELTS listening test through the use of lexical bundles[D]. Nottingham: Nottingham Trent University, 2022.

[32] ZHAO Y, ARYADOUST V. An automatized semantic analysis of two large-scale listening tests: A corpus-based study[J]. Language Testing, 2024, 42(3): 312-343. https://doi.org/10.1177/02655322241288598

[33] 许家金, 赵冲, 孙铭辰. 大语言模型的外语教学与研究应用[M]. 北京：外语教学与研究出版社, 2024: 40-50.

[34] ARYADOUST V, ZAKARIA A, JIA Y. Investigating the affordances of OpenAI's large language model in developing listening assessments[J]. Computers and Education: Artificial Intelligence, 2024, 6, 100204. https://doi.org/10.1016/j.caeai.2024.100204

[35] RUNGE A, ATTALI Y, LAFLAIR G T, et al. A generative AI-driven interactive listening assessment task[J]. Frontiers in Artificial Intelligence, 2024, 7. https://doi.org/10.3389/frai.2024.1474019

[36] WANG Y, MENG, Y. Optimizing distractor quality in a locally developed second language listening test: Integrating generative AI and psychometric methods[J]. Language Testing, 2025, 02655322251400375. https://doi.org/10.1177/02655322251400375

[37] 全国大学英语四、六级考试委员会. 全国大学英语四、六级考试大纲(2016 年修订版) [EB/OL]. (2016)[2025-10-12]. https://cet.neea.edu.cn/html1/folder/16113/1588-1.htm

[38] FAN J, FROST K, JIN Y. Local English testing in China's tertiary education: Contexts, policies, and practices[J]. Language Testing, 2022, 39(3): 453-473. https://doi.org/10.1177/02655322211070839

[39] NATION, I. How large a vocabulary is needed for reading and listening?[J]. The Canadian Modern Language Review, 2006, 63(1): 59-82. https://doi.org/10.3138/cmlr.63.1.59

[40] 金檀, 陆小飞等.阅读分级指难针[EB/OL]. (2025)[2025-12-28]. languagedata.net/tester

[41] 中华人民共和国教育部、国家语言文字工作委员会. 中国英语能力等级量表[S]. 上海：上海外语教育出版社, 2024: 8