ICS 35.240.99
L67

团　　体　　标　　准

T/ISC xxxx—202x

语音识别技术评测要求

Speech recognition technology assessment requirements

征求意见稿

2022.11

中　国　互　联　网　协　会　　发布

202x - xx - xx 发布　　202x - xx - xx 实施

目　次

前　言

本文件按照GB/T1.1-2020给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本部分由中国互联网协会提出并归口。

本部分起草单位：

本部分主要起草人：

语音识别技术评测要求

1 范围

本文件规定了连续语音识别评测测试集、评测方法、评测指标和评测报告的相关要求。

本文件适用于语音识别系统开发者、运营者及第三方机构对语音识别系统的连续语音识别能力进行测试和评估。

2 规范性引用文件

下列文件对于本文件的引用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 21023-2007 中文语音识别系统通用技术规范

3 术语、定义和缩略语

3.1 术语和定义

下列术语、定义和缩略语适用于本文件。

3.1.1 语音识别 speech recognition

将人类的声音信号转化为文字或者指令的过程。

3.1.2 语音识别系统 speech recognition system

具有语音识别功能的开发工具、软件、装置或应用。

3.1.3 连续语音识别 large vocabulary continuous speech recognition

指面向连续语音信号的自动语音识别，以区别于命令词语音识别。按照识别实时性要求，连续语音识别又分为流式和非流式两种类型。

3.1.4 删除错误 deletion error

用户所发语音在语音识别结果中没有出现的错误。

3.1.5 插入错误 insertion error

用户没有发的语音在识别结果中出现的错误。

3.1.6 替换错误 substitution error

用户所发语音被识别成其他语音的错误。

3.1.7 被测系统 system for testing

参加评测的语音识别系统。

3.1.8 测试系统 testing system

对被测系统进行评测的系统和体系。

3.1.9 测试语料 testing system

用于评测被测系统语音识别功能的音频集合。

3.2 缩略语

下列缩略语适用于本文件。

CER	Character Error Rate	字错误率
CCR	Character Correct Rate	字正确率
CSR	Continuous Speech Recognition	连续语音识别
WER	Word Error Rate	词错误率
WCR	Word Correct Rate	词正确率
MER	Mixed Error Rate	混合错误率

4 概述

本文件描述的评测要求主要围绕CSR开展，CSR之外的评测要求和指标不在本标准中进行描述。为保证语音识别系统评测的再现性，测试应尽量采用基于语音识别标准库的测试方法，无法采用基于语音识别标准库测试的，可采用基于现场口呼的测试方法。测试语料的设计与测试语音的录制应保证与实际使用场景的一致性，评测的结果以满足规范的评测报告形式给出。

5 测试集

5.1 概述

对用于语音识别评测的测试集，应建立语音识别标准库。标准库的建立应按照GB/T 21023的要求进行，通过专业录音麦克风在消音室环境下组织录制人员录制。本部分给出了测试语料设计要求和测试语音录制要求。

5.2 测试语料设计

测试语料应从词汇量覆盖、领域覆盖等方面加以设计。测试集文本上分成若干组，每组可以由若干人发音组成。设计要求如下：

a) 对于小词汇量每组测试集应包含所有词汇。
b) 对于中小词汇量每组测试集的合集应覆盖系统的所有词汇量。
c) 对于大中词汇量以上的测试集，每组测试集词汇的合集应考虑尽量多地覆盖系统词汇量。
d) 对在词汇、语法、语义等受到限制的连续语音，应充分考虑句型、词汇、语义等的覆盖性。
e) 对没有特别语言限制的连续语音，应从不同领域、不同应用场景考虑语料的选择，例如被测系统属于智能家电、娱乐直播、电话客服、公检法速记、智能教育、智能车载等不同应用领域，应在语料中考虑不同领域和应用下专有词汇、高频词汇的覆盖性。

5.3 测试语音录制

测试语音录制要求如下：

a) 说话人的选择应在符合系统对说话人限制的条件下，尽可能选择具有代表性和统计分布规律的发音人，特别是考虑不同口音、不同年龄、不同语速、不同教育背景、不同说话韵律等因素。
b) 测试语音的发音人至少为 30 个人以上，每人发音测试语料中的一组或多组语料，不同发音人尽量采用不同语料组。
c) 不同领域、不同应用场景的测试语音可根据各自特点设定环境背景（被测系统能正常工作的信噪比范围可能因应用场景的差异而不同）。
d) 测试语音的录制应与系统说明中的平台、采样率、输入通道等保持相对一致或接近，录音过程至少包括录音、标注和确认三个步骤，保证测试数据库的正确性。

6 评测方法

6.1 概述

连续语音识别的评测可采用基于语音识别标准库或基于现场口呼的方式进行。基于语音识别标准库的分为直接和间接两种测试方式，基于语音识别标准库的直接测试为录制语音数据的原声环境，间接测试和基于现场口呼的测试环境为混响环境。

6.2 基于语音识别标准库

基于语音识别标准库测试指采用录制的语音数据对被测系统进行直接或间接的测试，被测系统应至少满足其中一种测试方式。

a) 直接测试指利用被测系统带有的应用程序输入/输出接口，直接利用语音识别标准库中的语音文件进行测试；
b) 间接测试指评测系统利用高保真回放设备把语音识别标准库中的语音通过双方认可的方式输出到被测系统中。

6.3 基于现场口呼

现场口呼测试除了满足5.2和5.3的要求外，还需对现场操作进行记录和评估。

a) 需要有两个以上识别结果记录者，记录被测系统对当前发音的输出结果，记录表应包括发音人、记录人、操作人、监督人、发音内容、语音识别结果等内容；
b) 全部发音者测试结束后，统一按照性能标准进行指标评估，评估至少有两个人以上参与。
c) 对于识别结果能以文件形式给出的，被测系统按照发音人还应给出文件形式的输出结果，以便自动评测。

7 评测指标

7.1 准确率指标

连续语音识别结果通常可以表示成字、词的序列。连续语音识别结果中的错误分为插入错误、删除错误与替换错误。英文的连续语音识别系统识别结果一般以词为单位。相应的错误率为词错误率（Word Error Rate: WER），类似的语种还有俄语、维语等。中文存在分词歧义的问题，因此一般统计字错误率（Character Error Rate: CER），类似的语种还有日语等。

a) 中文连续语音识别评测中，假设正确文本字数为M，删除错误字数D_C、插入错误字数I_C和替换错误字数S_C，定义以下性能指标：
替代错误率：$S_{ER}=（S_C/M）\times100\%$
插入错误率：$I_{ER}=（I_C/M）\times100\%$

删除错误率：D_{ER}=（D_C/M）×100%
字错误率：CER=（（S_C+I_C+D_C）/M）×100%
字准确率：CCR=100%-CER

b) 英文连续语音识别评测中，假设正确文本单词数为N，删除错误单词数D_W、插入错误单词数I_W和替换错误单词数S_W，定义以下性能指标：
替代错误率：S_{ER}=（S_W/N）×100%
插入错误率：I_{ER}=（I_W/N）×100%
删除错误率：D_{ER}=（D_W/N）×100%
词错误率：WER=（（S_W+I_W+D_W）/N）×100%
词准确率：WCR=100%-WER

c) 针对多语种混杂建模单元不同的情况（如中英文夹杂）。假设多语种混合的正确文本字数为M，单词数为N，删除错误字数D_C、插入错误字数I_C和替换错误字数S_C ，删除错误单词数D_W、插入错误单词数I_W和替换错误单词数S_W，定义以下性能指标：
混合错误率：MER=（（S_C+I_C+D_C+S_W+I_W+D_W）/（M+N））×100%

7.2 实时率指标

在系统的标准配置条件下，假设发音从T_s开始，发音结束时间为T_e，识别结束时间为T_r，则实时率=（T_r-T_e）/（T_e-T_s），实时率越小，语音识别的识别效率越高。离线识别的情况，可按照识别时间与音频时长之比计算。

7.3 配置指标

被测系统正常运行语音识别所需的基本计算机配置，如CPU、内存、网络、麦克风、A/D精度等要求，由被测系统提供方给出。

8 评测报告

语音识别评测后应提交标准评测报告。报告应由以下几部分构成

a) 对被测系统的完整描述；
   1) 被测系统所能处理的词汇量等级，参考 GB/T 21023 词汇量分类。
   2) 被测系统所能识别的说话人人群的具体限制及适用范围。
   3) 被测系统所属领域及应用场景相关说明，包括特定领域和应用场景的语料设计说明。
   4) 被测系统麦克风与说话人的距离限制，麦克风性能要求，支持的 A/D 转换精度和采样率等。
   5) 被测系统能正常工作的信噪比范围。
b) 按照 GB/T 21023-2007 语音识别标准库及规范，描述测试数据的语音属性、测试词汇以及测试说话人的选择及确定情况。
c) 按照第 7 章定义的指标，给出各测试语音识别结果的相关指标及平均识别指标。
d) 评测过程的情况记录，采用的测试方法及运行过程的流畅性。
e) 被测系统的配置情况。

附 录 A
（资料性附录）
真实业务语音的采集与标注

当语音录制无法满足各领域评测需求时，可通过对真实业务语音数据进行采集和标注来建立测试集。设计要求如下：

内容方面，测试集内容需要保证一定的词汇量覆盖和领域覆盖，常见领域要求示例如下：

a) 智能家电： 包含智能音箱、智能电视、扫地机器人、陪伴机器人、可视门铃、智能门锁、智能灯、智能空调、智能风扇、智能电饭煲，智能油烟机等智能唤醒和操控等场景，高频词汇包含“启动”，“打开”，“关闭”，“返回”，“确认”，“调大”，“调小”等；

b) 娱乐直播：包含游戏直播，带货直播，线上 KTV，语聊房，短/长视频等泛娱乐内容审核和语义理解等场景，涉及的高频词汇如“王者荣耀”，“和平精英”，“中路”，“打野”，“青铜”，“吃鸡”，“下单”，“关注”，“点赞”，“收藏”，“K 歌”，“老铁”，“YYDS”，“橱窗”，“爆单”，“转发”等；

c) 电话客服：包含电信运营商，保险跟金融公司，电商跟贸易，交通跟物流等主流音转字语音交互场景，涉及的高频词汇如“电信”，“移动”，“联调”，“人工客服”，“投诉”，“地址”，“卡号”，“密码”，“金额”，“成本”，“快递”，“送达”，“查询”，“评价”，“满意”，“保价”，“合同”等；

d) 公检法速记：包含公安局审问笔记，法院庭审记录等离线异步保密音转字场景。涉及高频词汇包含“犯罪”，“侵犯”，“未成年”，“公安局”，“检察院”，“起诉”，“诉讼”，“维持原判”，“二审判决”，“休庭”，“控诉”，“原告”，“被告”，“控辩双方”，“证人证词”，“法律”，“道德”，“刑法”，“缓期”，“剥夺”，“政治权利”等；

e) 智能教育：包含一对一&一对多在线或线下课堂，涉及 ASR 的场景主要集中在口语测评和跟读练习等场景，涉及高频词汇如“英语”，“打分”，“朗读”，“会话”，“评测”，“发音”，“练习”，“弹奏”，“清音”，“新概念”，“作文”，“语义”，“语法”，“名词”等；

f) 智能车载：包含车载影音，车载导航，智能座舱等语音交互或播报场景。涉及高频词包含“播放”，“搜索”，“天气”，“地址”，“堵车”，“加油站”，“广播”，“故事”等；

标注方面，标注方案可参考 GB/T 21023。此外，测试集必须为精标数据（至少两次人工审核），数据标注字准率不低于 98%，数据须进行脱敏处理，且需根据不同业务应用提供数据的领域、语种、口音程度、噪音程度等信息。

参 考 文 献